DOCUMENT RESUME

ED 461 258 FL 024 538

AUTHOR Thornton, Julie, Ed.
TITLE Language Aptitude Invitational Symposium Program Proceedings (Arlington, Virginia, September 25-27, 1994).
INSTITUTION Center for the Advancement of Language Learning, Arlington, VA.
PUB DATE 1994-09-00
NOTE 220p.; For selected individual papers, see FL 024 539-545.
PUB TYPE Collected Works - Proceedings (021)
EDRS PRICE MF01/PC09 Plus Postage.
DESCRIPTORS Academic Advising; Affective Objectives; Age Differences; Attitude Change; Change Strategies; *Cognitive Style; Communicative Competence (Languages); Computer Oriented Programs; Educational History; English (Second Language); Government Role; Information Processing; *Language Aptitude; Language Processing; Language Proficiency; Language Research; *Language Tests; Learning Strategies; Linguistic Theory; Listening Comprehension; *Measurement Techniques; Migrant Education; Models; Modern Languages; Oral Language; Personality Traits; Predictor Variables; Psycholinguistics; Psychometrics; Second Language Learning; *Second Languages; Standardized Tests; Statistical Analysis; Student Characteristics; Test Construction; Test Reliability; *Testing
IDENTIFIERS Defense Language Aptitude Battery; Modern Language Aptitude Test

ABSTRACT

The Language Aptitude Symposium papers include: "Prognostication and Language Aptitude Testing, 1925-62" (Bernard Spolsky); "You, the Government, and Language Aptitude" (Government Roundtable; abstract only); "Styles of Thinking and Learning" (Robert Sternberg); "Current Research in Measuring Listening" (Robert N. Bostrom) (abstract only); "A Study of the Modern language Aptitude Test for Predicting Learning Success and Advising Students" (Madeline E. Ehrman); "The Investigation of Oral Proficiency and Language Learning Strategies in a Migrant ESL Context" (Helen Lunt) (abstract only); "Effecting Changes in Affective Factors" (Christine A. Montgomery) (abstract only); "The Defense Language Aptitude Battery (DLAB): What Is It and How Well Does It Work?" (John A. Lett, Jr., John W. Thain) (abstract only); "Expanding the Definition of Language Aptitude: The Role of Personality Variables" (Ehrman) (abstract only); "Zero-Based Language Test Design or Where's the Test's Focus" (Pardee Lowe, Jr.) (abstract only); "Aptitude Tests: Conception and Design" (James R. Child); "Models of Language Ability: Some Practical Considerations from a European Perspective" (John H. A. L. de Jong) (abstract only); "Aptitude from an Information-Processing Perspective" (Barry McLaughlin); "Learner Characteristics in Second Language Acquisition" (J. M. O'Malley, Anna Uhl Chamot) (abstract only); "Improving the Measurement of Language Aptitude: A Psychometric Analysis of the Defense Language Aptitude Battery" (Frances E. O'Mara, Thain) (abstract only); "Improving the Measurement of Language Aptitude: The Potential Contribution of L1 Measures" (Thain); "A Factor Analytic Study of Language Learning Strategy Use Older and Younger Adults" (Landes Holbrook, C. Eric Ott, Mary Lee Scott, Cheryl Brown) (abstract only); "Exploring Your Own Learning Style:

Reproductions supplied by EDRS are the best that can be made from the original document.

A Workshop" (Ehrman) (abstract only); "Psycholinguistic Issues in the Assessment of the Sub-Component of Language Abilities" (Brian MacWhinney); "Using Machine-Based Retrospective Correlation Data for Prospective Aptitude Assessment. Is Letting Computers Use the Past To Predict the Future Useful for Communicative Language Teaching?" (Frank Borchardt, Ellis Batten Page, Fred Jacome) (abstract only); and Test Theory and Language Learning Assessment" (Robert J. Mislevy). The conference program is appended. (MSE)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 461 258

1994
LANGUAGE APTITUDE
INVITATIONAL SYMPOSIUM

Center for the Advancement of Language Learning

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Julie Thornton
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

FL024538

Arlington, VA
September 25-27, 1994

BEST COPY AVAILABLE

LANGUAGE APTITUDE INVITATIONAL SYMPOSIUM

PROGRAM PROCEEDINGS

Sponsored by

Arlington, VA
September 25-27, 1994

Center for the Advancement of Language Learning
4040 N. Fairfax Drive, Suite 200
Arlington, VA 22203
Tel 703-312-5040 • Fax 703-528-6746

CALL 1994 LAIS Program Proceedings
Oversight: Peter A. Eddy
Editor: Julie A. Thornton
Cover Design: Katie Sprang

CALL 1994 Language Aptitude Invitational Symposium Working Group
Symposium Chair: Eduardo Cascallar
Program Chair: Julie Thornton
Program Manager: Katie Sprang
Invitations, Logistics, Registration & Materials: Claressa Strawn
Registration & Materials: Maria Vouras
Review Board: Marijke Cascallar, James Child, Kathleen Egan, Madeline Ehrman, John Lett, Jr., and Burt Weisman

Acknowledgments

The CALL 1994 Language Aptitude Invitational Symposium (LAIS) involved many people in many different roles. I would like to take this opportunity to acknowledge all of their efforts, without which the 1994 LAIS could not have taken place. It would be nearly impossible to describe all of their efforts and dedication here; however, perhaps the following short notes will help give credit where credit is due.

I would like to express my sincere appreciation to the individuals who presented papers at the LAIS. The plenary speakers drew from their varied expertise to provide complementary perspectives on language aptitude. These varied perspectives contributed much to the LAIS by providing new insights. Thanks also to the other presenters, both from government and academic organizations, for sharing the results of their research with others during the LAIS.

Members of the interagency CALL Research & Development Board played a vital role in refining the purpose of the LAIS, selecting plenary speakers, reviewing presentation proposals, and making presentations about their own organizations' history and interest in language aptitude. R&D Board members included the following individuals: Marijke I. Cascallar (FBI), James R. Child (Department of Defense), Kathleen Egan (Office of Research and Development), Madeline E. Ehrman (FSI), John A. Lett, Jr. (DLIFLC), and Burt Weisman (DIA). I thank them for the considerable amounts of personal and professional time they devoted to activities before, during, and after the LAIS.

Many thanks also to the CALL Executive Committee for its consistent support for the LAIS. The CALL Executive Committee consists of representatives of the following organizations: Center for the Advancement of Language Learning, Central Intelligence Agency, Defense Intelligence Agency, Defense Language Institute, Federal Bureau of Investigation, Foreign Service Institute, and National Security Agency. I also express our appreciation to the management personnel, testing program managers, language teachers and linguists of the organizations above who participated in the LAIS or who permitted their personnel to participate.

CALL staff and consultants, including Betty Kilgore, Candice Hunt, Earl Rickerson, and every member of the Programs staff helped to make the LAIS a success. Eduardo C. Cascallar was Symposium Chair, and Katie Sprang was my Program Co-Chair. Claressa Strawn was in charge of logistics, registration, and materials. Maria Vouras also assisted with registration and materials for the LAIS. I thank them for their hard work. Also, I thank Eva Taylor (FBI) for providing additional on-site registration assistance.

Julie Thornton

CONTENTS

(In the order in which the papers were presented.)

Introduction
Julie A. Thornton
Full Text of Paper 1

Prognostication and Language Aptitude Testing, 1925-62
Bernard Spolsky
Abstract 3
Full Text of Paper 26

You, the Government, and Language Aptitude
Government Roundtable Session
Abstract 4
Full Text of Paper ..... not available

Styles of Thinking and Learning
Robert Sternberg
Abstract 5
Full Text of Paper 51

Current Research in Measuring Listening
Robert N. Bostrom
Abstract 6
Full Text of Paper ..... not available

A Study of the Modern Language Aptitude Test for Predicting Learning Success and Advising Students[1]
Madeline E. Ehrman
Abstract 8
Full Text of Paper 74

The Investigation of Oral Proficiency and Language Learning Strategies in a Migrant ESL Context
Helen Lunt
Abstract 9
Full Text of Paper ..... not available

Effecting Changes in Affective Factors
Christine A. Montgomery
Abstract 10
Full Text of Paper ..... not available

[1] Original Title: "Is the Modern Language Aptitude Test Still Useful for Communicative Language Teaching?"

The Defense Language Aptitude Battery (DLAB): What is it and how well does it work?
John A. Lett, Jr. and John W. Thain
Abstract 11
Full Text of Paper ..... not available

Expanding the Definition of Language Aptitude: The Role of Personality Variables
Madeline Ehrman
Abstract 12
Full Text of Paper ..... not available

Zero-Based Language Test Design or Where's the Test's Focus
Pardee Lowe, Jr.
Abstract 13
Full Text of Paper ..... not available

Aptitude Tests: Conception and Design
James R. Child
Abstract 14
Full Text of Paper 100

Models of language ability: Some Practical Considerations from a European Perspective
John H.A.L. de Jong
Abstract 15
Full Text of Paper ..... not available

Aptitude from an Information-Processing Perspective
Barry McLaughlin
Abstract 16
Full Text of Paper 105

Learner Characteristics in Second Language Acquisition
J. M. O'Malley and Anna Uhl Chamot
Abstract 17
Full Text of Paper ..... not available

Improving the Measurement of Language Aptitude: A Psychometric Analysis of the Defense Language Aptitude Battery
Francis E. O'Mara and John W. Thain
Abstract 18
Full Text of Paper ..... not available

Improving the Measurement of Language Aptitude: The Potential Contribution of L1 Measures
John W. Thain
Abstract 19
Full Text of Paper 121

A Factor Analytic Study of Language Learning Strategy Use by Older and Younger Adults
Landes Holbrook, C. Eric Ott, Mary Lee Scott, Cheryl Brown
Abstract 20
Full Text of Paper ..... not available

Exploring your Own Learning Style: A Workshop
Madeline Ehrman
Abstract 21
Full Text of Paper ..... not available

Psycholinguistic Issues in the Assessment of the Sub-Components of Language Abilities
Brian MacWhinney
Abstract 22
Full Text of Paper 159

Using Machine-Based Retrospective Correlation Data for Prospective Aptitude Assessment. Is Letting Computers Use the Past to Predict the Future Useful for Communicative Language Teaching?
Frank Borchardt, Ellis Batten Page, and Fred Jacome
Abstract 23
Full Text of Paper ..... not available

Test Theory and Language Learning Assessment
Robert J. Mislevy
Abstract 24
Full Text of Paper 182

Appendix A: 1994 LAIS Program A-1

Introduction

The 1994 Language Aptitude Invitational Symposium (LAIS), and its proceedings, were sponsored by the Center for the Advancement of Language Learning (CALL). CALL was created in 1992 as a part of Congressional efforts to improve the foreign language capability of the US government. While its first aim is to strengthen language teaching and testing in federal organizations with needs for foreign languages, CALL is also a link to the academic and business language communities. It was with this in mind that the LAIS was planned and carried out. The conference was held near the end of September 1994 in Rosslyn, Virginia. Over 200 interested people from government, academia, and business gathered for the conference.

The US federal government has a critical interest in language aptitude since many agencies develop and use aptitude measures when hiring and/or assigning personnel and conduct research in language aptitude. The Interagency Language Roundtable Invitational Symposium on Language Aptitude Testing, the last language aptitude conference hosted by the federal language community, held in 1988, allowed government personnel involved in language aptitude research to exchange results and ideas with non-government researchers in language aptitude and other related fields. Selected papers from that symposium were published in *Language Aptitude Reconsidered*. The 1994 LAIS served the same function as the 1988 Symposium, updating government personnel on the latest work in language aptitude and other relevant areas, providing a forum for the exchange of new ideas and theories, and fostering greater interchange with academic researchers about the latest developments in the field.

The LAIS was sponsored by the government through CALL with the participation of researchers and practitioners both within and outside of government. The conference organizers took a rather broad definition of language aptitude, including individual differences research (including cognitive, motivational, affective, learning styles, strategies, etc.) along with other more traditional language aptitude topics. The LAIS provided an excellent opportunity to share new developments, both theoretical and applied, of benefit to the field of language aptitude and of interest to researchers and students in academic institutions as well as to government personnel. This volume of proceedings contains the abstract and biographical data for each of the papers presented at the LAIS. It does not contain all of the papers presented at the LAIS; only papers submitted for inclusion in ERIC are contained herein. This collection of papers provides a flavor of the LAIS and indicates the breadth of fields covered by the LAIS presenters, from the first paper by Bernard Spolsky on the historical development of aptitude testing to the last paper by Robert Mislevy on fruitful statistical analyses.

Abstracts

Prognostication and Language Aptitude Testing, 1925-62

Bernard Spolsky, Bar-Ilan University, Israel

During the 1930s, efforts were made in the US and elsewhere to develop prognostic tests that would justify decisions to exclude unqualified students from high school foreign language classes. In the US, after the second world war, government language programs supported research into the assessment of language aptitude to improve selection techniques. While an earlier study by one group of psychologists failed, later work by John B. Carroll and his colleagues led to the development of usable language aptitude tests, and contributed to the understanding of the nature of language aptitude.

Bernard Spolsky is Professor and Head of the Department of English at Bar-Ilan University, Israel. *This paper is based on one chapter in his book* Measured Words, *by Oxford University Press. Research for the book was carried out during a Mellon Fellowship sabbatical from Bar-Ilan University at the National Foreign Language Center in Washington, DC. This paper was published in* Language Testing *(vol. 12, no. 3, 1995) and appears in this collection with permission of the author and Edward Arnold, publisher of* Language Testing.

You, the Government, and Language Aptitude:

Roundtable Session

This roundtable session focussed LAIS participants' attention on the US government's practical uses for language aptitude tests as well as on current and future needs for such measures. Early in the LAIS, members of the CALL Research & Development Board gathered in a plenary information–sharing session to provide their perspective on language aptitude. They particularly underscored the very practical uses to which government agencies put language aptitude assessment: to provide decision-makers with a basis for making personnel assignments. At that session, each Board member spoke about his or her organization's procedures for language aptitude testing, described current needs in aptitude testing, and discussed future needs. The participants also shared their ideas about language aptitude and suggested likely directions for future research. They also discussed a number of important variables that might be incorporated into future aptitude assessments.

Styles of Thinking and Learning

Robert Sternberg, Yale University

Styles of thinking and learning are relevant to the understanding of foreign-language aptitude and for testing this aptitude. One particular theory of styles was emphasized: the theory of mental self-government. An outline of this theory and examples of each thinking and learning style were provided. In this theory, individuals differ as to their preferred thinking style, which Sternberg defines as preferred modes of thinking or of using one's abilities. A distinction was drawn between how well one thinks (relating this concept to ability) and how one thinks (relating this concept to style). It was stated that styles are often confused with abilities, so that students or others are thought to be incompetent not because they are lacking abilities, but because their styles of thinking do not match those of the people doing the assessment. Further, style is not an ability; rather, it is the way we use the abilities we have. Examples were provided of people who succeeded in different areas of their life by matching their style to their individual abilities. This model has been applied to career counseling and personal development. In conclusion, there was a call for teaching and testing to be done in such a way as to benefit individuals of all styles rather than to advantage individuals with one particular style.

Robert Sternberg is IBM Professor of Psychology and Education in the Department of Psychology at Yale University and current editor of the Psychological Bulletin. *His most recent book is titled* Defying the Crowd: Cultivating Creativity in a Culture of Conformity. *This paper was published in* Language Testing *(vol. 12, no. 3, 1995) and appears in this collection with permission of the author and Edward Arnold, publisher of* Language Testing.

Current Research in Measuring "Listening"

Robert N. Bostrom, University of Kentucky

Communication, for most of us, involves the exchange of messages, and usually is accomplished through speaking, listening, reading, and writing. When we say that we are studying communication, most of us mean that we are examining ways in which messages are designed, organized, mediated, or evaluated. But to assume that messages are received, processed, and retained in approximately the same form as the sender intended may be entirely unwarranted. Even very simple messages are easily distorted. Sometimes this distortion is caused by problems in attitude, in motivation, or in physical settings. But individual differences in receiving ability still account for large differences in communication effectiveness. Measuring the manner in which persons differ in this respect may be of great practical value.

Listening is probably the most common communication activity. In a much-cited study, Rankin (1929) asked persons to report how much time they spent in various types of communication. They reported that they listen 45 percent of the time, spoke 30 percent of the time, read 16 percent, and wrote 9 percent. In a more recent study, Klemmer and Snyder (1972) studied the communicative activity of technical persons. These persons spent 68 percent of their day in communication activity, and of that time, 62 percent was talking face-to-face. Klemmer and Snyder did not distinguish between speaking and listening, but it seems safe to assume that at least half of the face-to-face activity was listening. Brown (1982) estimates that executives in a modern corporation spend at least 60 percent of their day listening. To say that listening is an essential communication skill is to risk restating the obvious.

Research in listening focuses on decoding vocal messages and has usually used memory models as an exemplar (Loftus & Loftus, 1976; Collins & Quillian, 1972; Kintsch, 1980; McCloskey, 1980). Others concentrated on semantic memory (Baddely & Dale, 1968; Kintsch and Busche, 1969; Squire, 1986). Short-term processes are an important and often overlooked aspect of listening measurement (Schulman, 1972; Pelligrino, Siegel, & Dhawan, 1975; Monsell, 1984). The measurement of listening therefore involves several components: short-term listening, interpretive listening, and lecture listening. These aspects of listening have been the object of a good deal of research in the last few years (Bostrom & Waldhart, 1980; Bostrom & Waldhart, 1988, Bostrom, 1990). Initial findings support the separability of these component abilities, and of the different ability levels measured, the short-term measures seem to be the most valid in the way they predict other characteristics, especially success in organizational life (Bostrom, 1990; Sypher, Bostrom & Seibert, 1989; Alexander, Penley, & Jernigan, 1992). In other words, those who are skilled in short-term listening may or may not be skilled in lecture listening, which seems to be very closely related to common definitions of intelligence (Bostrom & Waldhart, 1988; Kelly, 1965, 1967).

Interpretive listening (vocalic decoding) may hold the most promise for current research endeavors. In one study using a standardized vocalic listening task, a very large sample of college students and adults only identified correct answers 55 percent of the time (Bostrom, 1990, p. 22). In other words, almost half of the time, people misinterpret vocalic signals. And while this kind of information access is universally considered to be of great importance, no one seems to have a clue as to how it should be improved (Samuels, 1987, p. 395). Research indicates that improvements in interpretive listening can be accomplished with training procedures, such as

sensitivity training, role playing, and the like (Wolvin & Coakely, 1985). Interpreting the underlying affect implied in spoken messages may involve personal schemata (Fitch-Houser, 1990), constructs (Crockett, 1989), or cultural literacy (Hirsch, 1987). These changes, however, are changes in attitude, awareness, or knowledge, not changes in basic ability. Changes in basic ability are much more difficult.

A substantive body of research clearly indicates that these interactions are also strongly affected by an individual's ability to decode the nonverbal cues present in the exchange (Archer & Akert, 1977; Buck, 1980, 1983; Burns & Beier, 1973; Ekman & Friesen, 1969; Furnham, Trevethan, & Gaskell, 1981; Mehrabian & Weiner, 1967; Zuckerman & Larrence, 1979). When nonverbal signals contradict the verbal ones, individuals typically accept the nonverbal as a more valid expression of the true feelings of the person with whom they are interacting (Burgoon, 1985; Leathers, 1979). Most investigations of nonverbal cues center on visual displays, such as facial expression, posture, and the like. Others have investigated vocalic messages, such as pitch, intonation, and inflection. Visual cues have been shown to be of greater influence than the vocalic ones in most situations. However, some studies show that vocalic cues are of more use in detecting deception than visual ones (Littlepage & Pineault, 1981; Streeter et al., 1977); apparently the traditional nonverbal categories are too simple (Keely-Dyreson, Burgoon, & Bailey, 1991). Research indicates that visual cues are decoded with much greater accuracy and that the ability to decode vocalic messages is not nearly as good as most people suppose. Circumstances may preclude the inspection of facial expression and other body movements.

First a defensible typological system of vocalics needs to be developed, and then compared with potential message systems inherent in visual signals. Measurement of these characteristics holds a good deal of promise for the prediction of success in most organizational settings.

Robert N. Bostrom (Ph.D., Iowa) is Professor of Communication at the University of Kentucky, Lexington. He is the author of a number of books, including Listening Behavior *(Guilford). He served as editor of the* ICA Communication Yearbooks, *and is the author of over fifty research articles and over a hundred convention papers. He developed the Kentucky Comprehensive Listening Test—a widely used research instrument. He was also a principal contributor to the National Teacher Examination's test of listening.*

A Study of the Modern Language Aptitude Test for Predicting Learning Success and Advising Students

Madeline E. Ehrman, National Foreign Affairs Training Center, Department of State

The Modern Language Aptitude Test (MLAT) was part of a project examining biographical, motivational, attitudinal, personality, and cognitive aptitude variables among a total of 1,000 adult students preparing for overseas assignments at the Foreign Service Institute (with various smaller Ns for subsamples completing different instruments). Data were analyzed using correlation, ANOVA, chi square, and multiple regression as appropriate to the data and the research questions. The MLAT proved the best of the available predictors of language learning success. As a part of an effort to expand the concept of language learning aptitude beyond the strictly cognitive, this study related the MLAT not only to end of training proficiency outcomes but also to personality disposition, using both overall correlational data and information on extremely strong and weak learners. The MLAT has been found to be about equivalent in its current predictive power (based on correlation) to the time it was developed. In addition, it was found to be especially powerful at the extremes of performance as measured by speaking proficiency. Also intriguing are the links between high scores on the MLAT and various other individual difference characteristics, including personality variables. Qualitative findings from use of the MLAT part scores in student counseling activities are also described, suggesting utility for the well-established instrument beyond prediction of learning success.

Madeline Ehrman is Director of the Research, Evaluation, and Development Division in the School of Language Studies at the US Department of State's Foreign Service Institute (FSI). She holds advanced degrees in Linguistics with a Ph.D. in Clinical Psychology. Her research emphasizes the role of personality factors in language learning, and she applies her findings to student consultation services available at FSI. Her latest book is titled Understanding Language Learning Difficulties *(Sage Publications); another,* Interpersonal and Group Dynamics in the Second Language Classrom, *co-authored with Zoltan Dornyei, is in press (Sage Publications). In recent years, she has published and spoken internationally on the subject of individual differences in adult language learning.*

The Investigation of Oral Proficiency and Language Learning Strategies in a Migrant ESL Context.

Helen Lunt, University of Melbourne

Some early findings were reported from a project examining the oral proficiency of subjects and their reported use of language learning strategies. A larger research project is planned to establish a relationship between learner strategies, language aptitude, and second language proficiency.

It has recently been suggested that language learning aptitude is a much wider concept than Carroll's four-factor theory. The possibility has been raised that language learning styles and strategies may be possible components, or correlates, of language learning aptitude and thus predictors of language proficiency (Oxford 1990). The project explored such links to establish a relationship between the factors of second language proficiency, learner strategies, and language aptitude.

In this initial stage of the research, 200 subjects took an oral interaction test in live interview and language laboratory formats. Their oral proficiency, demonstrated by performance on the test, and their language learning strategies, as identified by responses to a questionnaire, were examined with reference to age, sex, and native language.

Results indicating the relationship between oral proficiency and use of learner strategies, with reference to age, sex, and first language background, were presented in the form of correlations using the SPSS statistical package.

The next stage of this project will involve the testing of subjects' language aptitude. Thus it is hoped that the nature of the relationship between second language aptitude, proficiency, and learner strategies will be established, and that such knowledge will be of benefit to language teachers and learners in the design of placement and instructional curricula.

Helen Lunt is a Ph.D. candidate in the Department of Applied Linguistics and Language Studies at the University of Melbourne. She is particularly interested in the individual differences of second language learners.

Effecting Changes in Affective Factors

Christine A. Montgomery, Language Systems, Inc.

The SLA literature is replete with discussions of the effect of individual differences in language learning aptitude resulting from affective factors involving language learning situations. To oversimplify, individuals who have a history of successful experiences in language learning situations will—other things being equal—tend to exhibit a higher aptitude for language learning than individuals who have a history of unsuccessful experiences. For the latter, the classroom is a tribunal. The individual is the defendant, who is likely to be found guilty of making some gross error by all the others present if he dares to speak, and therefore remains silent as long as possible.

Without entering into the possible underlying socioeconomic factors that may lead to a generalized negative affective toward any learning situation, if we focus on the language learning situation and attempt to remove the stressful factors, it seems reasonable to assume that the negative affect could be reversed, or at least, reduced significantly, such that interference with language learning aptitude could be lessened. Even for individuals with a background of relative success, removal of the stress of "performing" in a language learning situation and risking failure while "on stage" should achieve improvement in language learning aptitude.

Innovative CALL technology that can potentially achieve this goal is now realizable, through utilization of systems integrating speech and natural language processing. Our group is currently engaged in the second phase of such a project, which will provide speaker independent, continuous speech translation from English to Spanish, Arabic, and Russian, and from these languages to English. The resulting system has the potential to provide a language learning environment equipped with an indefatigable, non-judgmental, speaker of the target language, who can engage the student in conversation, understand the student's response, provide feedback if desired, and answer the student's questions about word formation, grammar, or other semantic and pragmatic relations within its competence. Two sets of issues related to the system, including system issues, such as the tuning of the speech recognizer, and experimental design issues, such as the representativeness of the subjects for extrapolating results to larger populations of language learners, still remain to be addressed.

Christine A. Montgomery, who is LSI's president and founder, has more than 25 years of research experience in linguistics and natural language processing systems, with a special focus on language understanding and translation. She received her B.S. (French) and M.S. (Linguistics) degrees from Georgetown University, and her Ph.D. from UCLA (Anthropological Linguistics). She has conducted linguistic research on the Sebei language of Uganda, and she was involved in designing language training courses in Russian and French. Her experience also includes other European and African languages.

The Defense Language Aptitude Battery (DLAB): What is It and How Well Does It Work?

John A. Lett, Jr., Defense Language Institute Foreign Language Center
John W. Thain, Defense Language Institute Foreign Language Center

Each year, thousands of individuals are administered the Defense Language Aptitude Battery (DLAB) as part of their screening for possible training and service as military linguists. Of those who take the test, about one in ten actually receive foreign language training at the Defense Language Institute (DLI). This session introduced the DLAB and the context in which it is utilized to LAIS participants.

The session began with a brief discussion of the military recruiting and assignment system in which the DLAB is used as a major component in the multi-phase program that identifies potential military linguists. It was shown that the DLAB adds substantial and statistically significant variance to the prediction of language learning success, above and beyond that which is contributed by the general aptitude measures which are taken by all potential enlistees in the military services. Data was drawn from several studies, including a large-*n* longitudinal study know as the Language Skill Change Project (LSCP). Having established the general patterns of the DLAB's use, presenters discussed the nature of the DLAB's components. Developed in 1976, the DLAB is composed of 119 scorable items in seven principal sections: background data, voice stress pattern recognition, four deductive grammar sections, and a rule-inference section. These sections were described in sufficient detail to give participants a general understanding of the item types involved. The session concluded with a discussion of the observed value of the DLAB as it has been used over almost two decades. Correlation coefficients between DLAB scores and various measures of language training outcomes were presented. Both general trends and intriguing anomalies were also presented.

John A. Lett, Jr. is Director of Research and Analysis at the Defense Language Institute Foreign Language Center, where he manages and directs the DLI research program. His division conducts and coordinates research performed at or for DLI personnel or others. Major division projects have addressed topics such as language proficiency change over time, aptitude and other predictors of language learning outcomes, awareness and optimal use of learning styles and strategies, and appropriate uses of educational technology in FL learning.

John W. Thain (M.A., UCLA) is an Educational Researcher at the Defense Language Institute. He was formerly involved in the foreign language proficiency testing program at the DLI. His current research interests include testing the listening comprehension of native English speakers, ethnography as a research tool in foreign language classrooms, language aptitude testing (with particular emphasis on measures used to select DoD linguists), programs of instruction and counseling in foreign language learning strategies, the linguistic classification and typology of foreign languages, and programs for cross-training DoD linguists into new languages.

Expanding the Definition of Language Aptitude: The Role of Personality Variables

Madeline Ehrman, National Foreign Affairs Training Center

Personality and learning style have been normally used for activities like student counseling and curriculum design. Nevertheless, it has been clear that individuals with certain learning styles (measured by a personality inventory, the Myers-Briggs Type Indicator (MBTI)) have been more comfortable with different methodologies. In terms of the MBTI, audio-lingual methodology, with its highly structured approach to teaching, is likely to appeal to sensing and judging types, whereas more communicative activities, with their high level of ambiguity, are more comfortable for intuitives and perceivers. All language learning requires students to cope with ambiguous input and incomplete understanding (since even the most highly structured methods have to expose students to real language sooner or later). Thus it seems possible that learning style characteristics reflecting tolerance of ambiguity would correlate with success in communicative and even naturalistic language learning. This presentation reports on research on a variety of individual differences to find the relationships among biographic data, cognitive aptitude, personality, learning strategies, and other learning styles variables, as well as to outcome data. The outcome data include both end-of-training proficiency scores in speaking and reading (based on oral interviews) and teacher ratings. (Tolerance of ambiguity is assessed through the MBTI and the Hartmann Boundary Questionnaire, which addresses relative permeability of ego boundaries.) There are implications for definitions of language aptitude that go beyond the strictly cognitive. The findings are also used to build a model of learning that goes from relatively deep in the personality (thickness or thinness of ego boundaries) through behavior to performance with four personality-based tracks much like learning styles applicable to learning and student counseling. An individual student will prefer to use of one or more of these tracks. The content of this paper is covered thoroughly in the following references:

Ehrman, M.E. (1993). Ego Boundaries Revisited: Toward a Model of Personality and Learning. In J.E. Alatis (Ed.), *Strategic Interaction and Language Acquisition: Theory, Practice, and Research*. Washington, DC: Georgetown University Press, pp. 331-362.

Ehrman, M.E. (1996). *Understanding Language Learning Difficulties*. Thousand Oaks, CA: Sage Publications.

Ehrman, M.E. (in press). *Ego Boundaries and Tolerance of Ambiguity in Second Language Learning*. In J. Arnold (Ed.), Affective Language Learning. New York: Cambridge.

Madeline Ehrman is Director of the Research, Evaluation, and Development Division in the School of Language Studies at the US Department of State's Foreign Service Institute (FSI). She holds advanced degrees in Linguistics with a Ph.D. in Clinical Psychology. Her research emphasizes the role of personality factors in language learning, and she applies her findings to student consultation services available at FSI. Her latest book is titled Understanding Language Learning Difficulties *(Sage Publications); another,* Interpersonal and Group Dynamics in the Second Language Classrom, *co-authored with Zoltan Dornyei, is in press (Sage Publications). In recent years, she has published and spoken internationally on the subject of individual differences in adult language learning.*

Zero-Based Language Test Design or Where's the Test's Focus

Pardee Lowe, Jr., Federal Language Training Laboratory

Language aptitude test design should be viewed afresh. The limited construct(s) of language aptitude and disagreement over which components contribute to such constructs are two shortcomings of current language aptitude tests. For many researchers, the current conception of language aptitude now includes motivation, learning strategies, and teaching styles. The insufficiencies of current aptitude tests are further described in the light of uses to which the US Government puts language aptitude test results. To this end, the following seven questions need to be asked of current and future aptitude tests with supporting background information.

- Can an aptitude test tell us that somone can learn a second or another language?
- Do languages have personalities, and is it possible to match or mismatch a language to a person? If so, what are the potential effects of such a mismatch?
- How difficult a language can the individual handle?
- What language (types) can the examinee most likely master?
- In what skill modality or modalities (speaking, listening, reading, writing) will the examinee most likely excel?
- In the case of a person who is to be selected for participation in a language training course, how well will the examinee do given the stated course goal(s)?
- How far can a person *ultimately* go in learning a given (type) language?

Government users have a vital interest in finding answers to these questions. Government researchers should attempt to apply the results of their work on language aptitude to answering these questions for those who use aptitude tests. Certain models of test design could conceivably lead us further towards answering these questions. The designs included are rather abstract and reified compared to what real-world test design has heretofore encompassed. They are purposefully kept simple to show the pure possibilities, yet may be easily expanded to conform to the necessity of real-world language aptitude test design. Among other points, the paper draws attention to a continuum of test design possibilities ranging from *One Size Fits All* approach at one extreme to the *Chinese Menu: One from Column A and One from Column B* approach at the other. Drawing on the previous sections of the paper outlining the seven questions, the insufficiencies of current language aptitude tests, and suggested future test design models, the paper concluded by focussing test design beyond our former limited conception of the construct and suggested possible fruitful approaches for the future.

Pardee Lowe, Jr., is past chair of the testing committee of the Interagency Language Roundtable (ILR), with experience in running such training programs, with testing design in general, and with the ALAT, the MLAT, the Pimsleur Aptitude Battery, and VORD in particular, suggested to him long ago the need for more reliable predictors. He is concerned with expanding the construct "language aptitude" and with more specific questions deriving from the practical application of such test results in government.

Aptitude Tests: Conception and Design

James Child, National Security Agency

This paper is concerned with the cognitive aspects of language aptitude testing as they affect prospective government language learners. Differences in skill modalities are considered (speaking, listening, reading), as are the levels of attainment that aptitude measures would ideally predict in the respective skills. Mention is made of the advantages previously studied languages offer in supplementing or serving as surrogates for aptitude tests, especially if the target (third) language is typologically similar to the language(s) already learned. Finally, some comments are offered on the strengths and weaknesses of existing models.

James R. Child, (M.A., University of Pennsylvania) has taught several languages, including Czech, Indonesian, Portuguese, and Turkish. He has over 40 years experience in language testing for the Department of Defense, including work on test design, the ALAT, the MLAT, and VORD. Since 1970, he has directed his organization's work-related language testing program. He is concerned with the philosophical underpinnings of language testing.

Models of Language Ability: Some Practical Considerations From a European Perspective

John H.A.L. de Jong, CiTO (National Institute for Educational Measurement), The Netherlands

Though CiTo's work has focused exclusively on ability testing, it was hoped that a professional's view of current developments in language testing with an emphasis on European efforts would characterize some of the variables involved and thus contribute to this discussion on language aptitude testing. Language aptitude is a hard nut to crack because it requires that we predict the probability for the future development of specific psychological traits. The observation, let alone the measurement, of such developmental change is usually difficult.

Current developments in Europe on issues of language testing were discussed in the first part of this talk as background. In contrast to the ACTFL guidelines used in the US, which have been around for a long time, European researchers have been building a shared language learning and testing framework, but are not yet done. In Europe, a rather chaotic approach is taken to organizing education, including language training. There is general agreement that there should be testing within the school system, but these exams are administered at arbitrary points in the student's education. Each European country has approached this work on its own, and a more structured view is needed for articulating the curriculum across levels, across languages, and across national borders. The LangCred project was also discussed. The first half name of this project derives from the word *language*. This group is attempting to make sense of the variety of certificates and diplomas to create a common currency in language ability certificates in all EU countries. For the students, the second half of the name stands for the word *credit*, in that is it a way for students to get credit for work they have done. For employers the second half of the project name stands for the word *credibility*, so they know how much they can rely on a certificate when they are making personnel decisions. The researcher defined six rating levels for each of two important aspects of language functioning—language ability and professional/vocational operations—and then tried to rate all existing diplomas and certificates against those levels.

In the second part of this paper, a general model of language learning from a measurement point of view was presented taking these developments into account. It is important to include variables related to the person, the task, and the situation. Most tests take only the task into account, and consider individual and situational differences as external variables. Once the results are collected, they are usually analyzed using correlations or exploratory factor analyses. Correlation approaches seem to be insufficient for measuring language aptitude. This is shown in the results of a number of studies that report that the correlation between language aptitude scores and actual performance in language tasks is moderate at most and low in general. This correlation approach should only be the first step. Additional differential approaches would provide a broader view of the variables at work.

John H.A.L. de Jong is the Director of the Language Testing Unit at CiTO (National Institute for Educational Measurement) for the Netherlands.

Aptitude From an Information-Processing Perspective

Barry McLaughlin, University of California, Santa Cruz

This paper outlines an information-processing approach to language learning, fits aptitude into that approach, and discusses what may be one aspect of aptitude—working memory. The process of learning includes two processes that make heavy use of working memory: automatization and restructuring. At first, learners must make a conscious effort to remember and apply a new concept early on, but later can apply the same concept without that conscious effort. Thus, the initial stages of learning involve the slow development of skills and the gradual elimination of errors as the learner attempts to automatize aspects of performance. About restructuring, individual differences in language learning aptitude are suspected to be the result, in large measure, of the joint function of availability of knowledge about the target language and the speed and efficiency of working memory—which affects the extent to which the individual succeeds in generalizing and altering the cognitive data required at various [language] processing stages. That is, in L2 learning working memory relates to the degree to which individuals can more flexibly and consistently restructure and reconfigure linguistic representations.

Barry McLaughlin is a Professor in the Program in Experimental Psychology at the University of California, Santa Cruz, where he is also Director of the Bilingual Research Group and Co-Director of the National Center of Research on Cultural Diversity and Second Language Learning. *This paper was published in* Language Testing *(vol. 12, no. 3, 1995) and appears in this collection with permission of the author and Edward Arnold, publisher of* Language Testing.

Learner Characteristics in Second Language Acquisition

J. M. O'Malley, Prince William County Public Schools
Anna Uhl Chamot, Director, Language Research Projects, Georgetown University

This paper (1) examined learner characteristics and aptitudes in second language acquisition, (2) advanced a cognitive-theoretical framework for understanding these characteristics, and (3) applied the theoretical framework in analyzing the influence of these characteristics on individual differences in learning. The range of learner characteristics that has been considered as influencing second language acquisition is extremely broad and includes exhaustive listings of characteristics such as age, gender, attitude, aptitude, personality, and cognitive style. These listings are typically accompanied by an analysis of why each characteristic influences second language acquisition with correlations between these characteristics and learning outcomes. While interesting, these analyses have not provided an integrated view of learner characteristics that influence learning outcomes and which are responsive to instruction. No one has yet advanced a central theoretical position at the onset that can be used to select learner characteristics, indicate why these particular characteristics are expected to influence language learning outcomes, and then examine research evidence to suggest their level and type of influence.

A cognitive-theoretical view of second language acquisition was detailed. Learner characteristics that according to the theory should influence learning outcomes were specified. Other representative learner characteristics suggested by using the theory that may be important for instruction were also examined. Promising research methods analyzing these learner characteristics and the instructional implications of the theory for these characteristics are identified. Learner characteristics have significance in second language acquisition because of their relationship to individual differences in the rate of learning and level of proficiency individuals attain. That is, variables such as age, motivation, learning style, or aptitude may influence the ways in which individuals go about learning a second language, their rate of learning, or their ultimate proficiency in using the language effectively. Learner characteristics are also of interest to researchers and theorists to improve our understanding of these variables and the ways in which they are interrelated in conceptual models of second language acquisition.

J. M. O'Malley is presently Supervisor of Assessment and Evaluation in Prince William County Public Schools. His interests are in alternative assessment, cognitive theory, and applications of learning strategies in second language instruction. He has conducted research on learning strategies in second language acquisition and published extensively on the research and on an instructional model he co-developed with Dr. Chamot.

Anna Uhl Chamot is Director of the Language Research Projects and Adjunct Professor in the Department of Linguistics at Georgetown University. Her interests are in second language acquisition and in staff development, materials development, and applications of learning strategies to second language instruction. As an extension of a long line of research, she is currently conducting studies on instruction and strategic approaches to foreign language learning.

Improving the Measurement of Language Aptitude: A Psychometric Analysis of the Defense Language Aptitude Battery

Francis E. O'Mara, PRC, Inc.
John W. Thain, Defense Language Institute Foreign Language Center

The Defense Language Aptitude Battery (DLAB) is administered to thousands of individuals each year as part of their screening for training as military linguists. Of those who take the test, about one in ten actually receive foreign language training. The DLAB is used to select which ten percent of the potential students have the highest aptitude for learning a foreign language as well as suggesting the kind of language that each can successfully undertake.

Developed in 1976, the DLAB has served competently for almost two decades. Periodic assessments by DLI of how well DLAB scores predict success in DLI training have consistently shown significant results. These observations were reinforced by the findings of the Language Skill Change Project, which showed that DLAB scores were among the most potent predictors of DLI training outcomes from a wide variety of cognitive, affective, and background variables. In 1990, DLI undertook a program of research intended to identify ways to improve the selection and assignment of future military linguists. Increasing the predictive power of the existing DLAB by adding language-specific and skill-specific prediction capabilities was the objective.

An extensive item-level analysis of the existing DLAB was conducted to reveal opportunities for short-term improvement in the test as well as to provide information on language learning aptitude useful in future efforts to design a replacement test. The analysis proceeded along three lines of inquiry: (1) a classical psychometric item analysis, to determine which items and measurement features of the DLAB contribute to or detract from its demonstrated validity; (2) an exploration of the potential to improve the DLAB's predictive power by considering examinee score profiles by sections rather than using a single score; and (3) an examination of the test's dimensionality through a series of factor analyses to identify the components of aptitude measured by the test and determine how each was involved in the prediction of DLI student performance. The session described the data sources and sample characteristics, summarized the analysis methodology and results, and indicated how results were used in modifying the current DLAB.

Francis E. O'Mara (Ph.D., University of Delaware) serves as a Technical Director with PRC, Inc. in McLean, Virginia where he directs research and development projects on issues pertaining to personnel and training. Over the last ten years, he has conducted research efforts concerning the nature and measurement of foreign language aptitude, and the acquisition and long-term retention of foreign language skills.

John W. Thain (M.A., UCLA) is an Educational Researcher at the Defense Language Institute. He was formerly involved in the foreign language proficiency testing program at DLI. His current research interests include listening comprehension testing, ethnography as a research tool in foreign language classrooms, language aptitude testing of DoD linguists, language learning strategies instruction and counseling, foreign language classification and typology, and cross-training programs for DoD linguists.

Improving the Measurement of Language Aptitude: The Potential Contribution of L1 Measures

John W. Thain, Defense Language Institute Foreign Language Center

Two major studies at the Defense Language Institute investigated the contribution of several variables to prediction of post–language–training proficiency. These predictor variables included (1) scores on a general vocational aptitude battery and a language aptitude battery, both used to screen potential students; (2) scores on other cognitive measures, not used in the screening process; and (3) scores and ratings on measures of student motivation, anxiety, and use of learning strategies. Follow-ups to these studies lend impetus to an effort to add supplemental predictors to the language aptitude battery used for selection of students to attend DLI. These follow-on efforts addressed the potential value of adding certain types of native language competency measures to the existing aptitude battery.

The first effort consisted of a review of the literature of currently used native-language listening tests. The review highlighted the differences between the serial processing inherent in listening comprehension as opposed to the parallel processing involved in reading comprehension. Two important factors influencing the difficulty of tasks in listening tests were identified as (1) the extent to which the examinee had the opportunity to rehearse the initial stimulus or to recode it for later use in performing the testing task, and (2) the extent to which the examinee had a preexisting mental set enabling him/her to apply an appropriate schema to select and organize the stimulus input as needed to perform the testing task. The review of the literature also investigated the role of pragmatic conversational inference in interpreting conversational discourse.

A second follow-on effort involved the review of tests of English grammar with particular focus on speeded tests in which the examinee's task was to identify grammatical errors in English sentences. In addition, during the course of this work, an extended comparison was made between (a) English grammar tests and (b) conventional subtests of grammatical skills that utilize an artificial language work sample. This session described the results of these follow-on efforts, discussed their implications for language aptitude test battery development, and generated collegial feedback from session attendees.

John W. Thain (M.A., UCLA) is an Educational Researcher at the Defense Language Institute (DLI). He was formerly involved in the foreign language proficiency testing program at DLI. His current research interests include listening comprehension testing, ethnography as a research tool in foreign language classroom, language aptitude testing of DoD linguists, language learning strategies instruction and counseling, foreign language classification and typology, and cross-training programs for DoD linguists.

A Factor Analytic Study of Language Learning Strategy Use by Older and Younger Adults

Landes Holbrook, Brigham Young University
C. Eric Ott, Missionary Training Center
Mary Lee Scott, Brigham Young University
Cheryl Brown, Brigham Young University

Since older learners have been largely ignored in second language acquisition research, they were chosen as the focus of a study on language learning strategy use. Subjects were 26 older adults (46-70 years old) and 235 younger adults (approximately 19–25 years old) learning a variety of second languages in an intensive eight-week course. Strategy use was assessed through a self-report questionnaire based on the one developed by Oxford (1990), which groups items as representing memory, cognitive, compensation, metacognitive, affective, and social strategies.

A factor analysis derived groupings of strategies different from those reported by previous factor analyses (e.g., Oxford & Nyikos, 1989). These groupings were used in analyses of variance comparing older and younger learners. In addition, a subset of data was also evaluated to determine how language learning strategy use was affected by gender and language proficiency. Possible reasons for the divergence from previous factor analyses were discussed. The relationships between factor scores and the variables of age, gender, language proficiency, and learning context were also examined. Implications of these results for teaching older and younger learners was also addressed. In addition, there was a discussion of the importance of considering a new class of learning strategies which emerged from this study and which appeared to be of great significance to the learners.

Landes Holbrook (M.A., BYU) is a teacher supervisor at the Brigham Young University English Language Center.

C. Eric Ott (Ph.D., BYU) is the Director of Research and Evaluation at the Missionary Training Center in Provo, Utah. Two of his current areas of interest are language learning strategies and task-based language learning.

Mary Lee Scott (Ph.D., UCLA) is an Assistant Professor in the Linguistics department at Brigham Young University, where she teaches courses in the TESL and Language Acquisition M.A. programs. She is interested in researching the language learning of older adults, issues in language testing, and use of language learning and communication strategies.

Cheryl Brown (Ph.D., UCLA) is Associate Dean of the College of Humanities at Brigham Young University and teaches courses in the TESL and Language Acquisition M.A. programs.

Exploring Your Own Learning Style: A Workshop

Madeline Ehrman, National Foreign Affairs Training Center

Within the broad range of factors that contribute to language aptitude are individual differences in learning style. This workshop was designed to help participants gain a sense of their own learning styles in one model. Various models were described, including sensory channels (visual, auditory, kinesthetic). However, the primary emphasis was on the Myers-Briggs Type Indicator (MBTI), which participants took during the workshop. The MBTI was used because it probably has the greatest theoretical depth of any learning styles model currently available.

The three-hour workshop was organized as follows:

Initial discussion of observed individual differences (15 min.)
Participants take the MBTI (30 min.)
Description of the meaning of the MBTI dimensions (50 min.)
Application activity (1 hour)
Discussion of applications to participants' settings (15 min.)

Participants learned about their own styles. The relationship of styles to language learning were also addressed. Participants had the opportunity to discuss the application of this model within their own organizations.

Madeline Ehrman is Director of the Research, Evaluation, and Development Division in the School of Language Studies at the US Department of State's Foreign Service Institute (FSI). She holds advanced degrees in Linguistics with a Ph.D. in Clinical Psychology. Her research emphasizes the role of personality factors in language learning, and she applies her findings to student consultation services available at FSI. Her latest book is titled Understanding Language Learning Difficulties *(Sage Publications); another,* Interpersonal and Group Dynamics in the Second Language Classrom, *co-authored with Zoltan Dornyei, is in press (Sage Publications). In recent years, she has published and spoken internationally on the subject of individual differences in adult language learning.*

Psycholinguistic Issues in the Assessment of the Sub-Components of Language Abilities

Brian MacWhinney, Carnegie-Mellon University

Drawing upon recent psychological and neurological research related to how individual differences might interact with learning a particular language, an attempt was made to show how psycholinguistic research and theory can help in the process of assigning military personnel to language training and to a given language. Using the Defense Language Institute's Defense Language Aptitude Battery (DLAB) as a point of departure, five categories of language difficulty were described, based on their degree of difference from English. Research on individual differences in language learning was reviewed to identify potential methods of improving the DLAB to measure individual traits which may allow a student to excel at certain languages or language skills. To make these assessments, the specific areas in which language presents learning difficulties include orthography, phonology, lexicon, morphosyntax, syntactic processing, language use, and language learning strategies. In each of these areas, there is a rich psycholinguistic literature that can be used as a basis for elaborating additional tests of learner abilities. As a practical matter, it may be useful to place increased emphasis on the prediction of L2 learning from L1 language skills. In addition, it will typically be easier for practical reasons to measure skill at learning and processing orthography and receptive phonology than on-line sentence structure and productive phonology. There is a need for more "fine-grained psycholinguistic measures of language learning skills as they are applied during the various stages of language instruction" if the goal to is to improve prediction of problems caused by individual differences, and the research described a number of areas in which tests could be developed that would provide information on the interaction of individual language ability and a particular language's difficulty level.

Brian MacWhinney is currently a Professor of Psychology at Carnegie Mellon University, where he serves as Director of the Child Language Data Exchange System Project. He has performed research and written extensively in the areas of psycholinguistics and first language acquisition. *This paper was published in* Language Testing *(vol. 12, no. 3, 1995) and appears in this collection with permission of the author and Edward Arnold, publisher of* Language Testing.

Using Machine-Based Retrospective Correlation Data for Prospective Aptitude Assessment. Is Letting Computers Use the Past to Predict the Future Useful for Communicative Language Teaching?

Frank Borchardt, Duke University
Ellis Batten Page, Duke University
Fred Jacome, Duke University

Introducing automation to the aptitude assessment process has been found to be problematic. The authors have performed research dealing with the role of the computer in what are traditionally considered hard-to-quantify areas of human activity. Machine-based evaluation of hard-to-quantify activities (such as the grading of English compositions written by high school students) can produce cross-validation results as high as .87 when compared to the performance of a single human judge. Furthermore, machine-based grading can be higher than the inter-rater reliability level of two human judges; i.e., it has been found that machine-based evaluation can accord with a human judge more often than two human judges accord with one another. In contrast to human raters, computers employ indirect criteria. In the case of written essay evaluation, direct or intrinsic variables of interest might include fluency, diction, style, organization, or logic. Measurable substitutes, called approximations, have been used to approximate these variables, such as essay length for fluency, variation in word length for diction, proportion of subordinating conjunctions and relatives for complexity of style. Application of the underlying principles of this research to aptitude testing would seem a highly plausible and labor-saving strategy.

Modeled on the English essay experiment, where criteria, both direct and indirect, intrinsic and approximate were established a priori and applied to testing data, an aptitude testing experiment was proposed. This experiment would include individuals with a demonstrably high degree of language aptitude. Data on these subjects, such as all available electronic performance, drill-and-practice, quizzes, tests, questionnaires, would be compiled into a comprehensive database. This database, which could be built using WinCALIS, which would then be analysed to identify all statistical correlatives. Later, these variables could be classified as intrinsic or approximate, to which a third, more remote and accidental category, called contingent could be added. Such results could be employed to identify, prospectively and with high probability, individuals especially well equipped with language aptitude.

Dr. Frank L. Borchardt (Ph.D., Johns Hopkins University) directs the project at Duke University which produced CALIS (Computer Assisted Language Instructional System) and WinCALLIS.

Dr. Ellis Batten Page was founding editor of the Educational Psychologist *for the American Psychological Association.*

Fred Jacome is Project Manager of the WinCALIS Project at the Humanities Computing Facility at Duke University.

Test Theory and Language Learning Assessment

Robert J. Mislevy, Educational Testing Service

Standard test theory is machinery for carrying out inference in a particular admixture of ideas from statistics, measurement, and psychology that coalesced in the first third of this century. Recent developments in cognitive and educational psychology, such as increased appreciation of the situated nature of learning and understanding, call for broader ranges of student models and types of data than those that are standard in testing today. We must specify how what we observe on the test is related to competence as we choose to conceive it, and construct a framework for carrying out inferences, i.e., making decisions about a test-taker's language based on the test results, within the framework we thus erect. At present, there is a growing need for testing that verifies performance within a given situation or context or in the completion of a complex task that requires the language learner to integrate a number of aspects of his or her learning. A broader view of testing to include these issues requires the clear definition of what is to be measured, new methods of eliciting evidence of what is to be measured, and methods of weighting the various sources of evidence for use in making decisions about test results. For the latter, it is possible to analyze statistically the evidence so that the mathematical models mirror the model of drawing inferences based on that evidence. A conceptualization of test theory is discussed which is meant to address issues of weight and coverage of evidence for statements framed in more recent educational/psychological paradigms. Using this technique, a number of examples, including inferences based on the ACTFL Guidelines, was presented.

Robert Mislevy is a Principal Research Scientist in the Model-Based Measurement Group in the Statistical and Psychometric Research and Services division at Educational Testing Service, Princeton. He is Past President of the Psychometric Society. *This paper was published in* Language Testing *(vol. 12, no. 3, 1995) and appears in this collection with permission of the author and Edward Arnold, publisher of* Language Testing.

Full Text of Papers

PROGNOSTICATION AND LANGUAGE APTITUDE TESTING -- 1925-62

Bernard Spolsky
Language Policy Research Center, Bar-Ilan University and the National Foreign Language Center, Johns Hopkins University

It is a signal honor to have been invited to address this Symposium, a further important milestone in the long-established collaboration between academic language testers and the government language teaching and testing establishment.[1] While the first interest in language aptitude came from the colleges and universities in the 1920s, the major developments in language aptitude testing in the 1960s were the result of government initiative, and it is most fitting that CALL should have taken the lead in this intended to continue the refinement of the field.

Language testing is a field that has long recognized its social and political significance. A hundred years before Foucault, in the brilliantly stimulating few pages he devoted to examinations, showed their disciplinary effect in providing 'un regard normalisateur, une surveillance qui permet de qualifier, de classer et de punir' (Foucault, 1975:186-7),[2] Henry Latham (1877) was already decrying the "encroaching power" of examinations that, he protested, was biasing education, blurring important distinctions between liberal and technical education, and narrowing the range of learning through forcing students to prepare by studying with crammers and in cramming schools.

Ironically, examinations had long been regarded as forces for good and a method of attaining to equal opportunity. The original Chinese system, that lasted two thousand years, was intended to recruit civil servants on the basis of their excellence rather than

---

[1]This paper was based on a chapter from my book, *Measured Words*, published by Oxford University Press in March 1995. Research on it was carried out while I was on sabbatical leave from Bar-Ilan University as a Mellon Fellow at the National Foreign Language Center. It has been revised to be the opening plenary paper at the 1994 Language Aptitude Invitational Symposium sponsored by the Center for the Advancement of Language Learning, held at Rosslyn, VA, from September 25-27, 1994. This is the version prepared for the meeting. An edited version has also been published in *Language Testing*, 12 (3) 321-340, 1995.

[2]'a normalizing gaze, a surveillance that makes it possible to qualify, to classify and to punish'.

their patronage, and it was this 'Chinese principle' that was used by Lord Macaulay to bolster arguments for using examinations for selecting cadets for admission to the India Civil Service that was one of the major reforms in nineteenth century England. The egalitarian potential of the public examination no doubt contributed to its importance in the United States after independence and in Revolutionary France, although clearly Napoleon saw its potential for centralized control.

It was concern with the fairness of powerful public examinations that led Edgeworth (1888) to call attention to their "unavoidable uncertainty." The new-type objective test was seen as a solution to this problem. Objective testing started to increase in Britain and the United States in the decade or so after the First World War, but only in America did it find an immediate public acceptance, as the testing business started to sweep American education in the late 1920s

Language testing was not immune to objectivisation. By 1930, the work of the Modern Language Study had demonstrated that the achievement test or examination could be a powerful tool for control over the language teaching process, and in the hands of the College Entrance Examination Board, the proficiency test or examination was developing into an equally effective way to maintain authority over the language qualifications of applicants for admission to universities or countries. Between the World Wars, these tests evolved steadily, with constant progress towards objectivization and industrialization that I have discussed elsewhere (Spolsky, 1995).

There remained another area of disquiet, the control of admission to the language learning class itself, and it is with this parallel development that this Symposium and this paper that opens it will deal. In the first half of the paper, I will describe attempts made in the U.S. between the two world wars to develop prognosis tests with the goal of ensuring that only qualified students would be allowed into high school language classes. In the second, I will describe two research programs, one a failure and the other a major success, to develop aptitude tests that would allow government agencies to select only appropriate candidates for expensive intensive language training.

There are two main points that this study will reveal. The first is that the level of success of the efforts was more a function of the resources made available to the task than of the state of knowledge or sophistication of the researchers. The pre-World War II enterprise of language teachers to control access to their classes was local and conducted with minimal funds; nonetheless, useful tests were developed and a general theoretical model of considerable sophistication was established.[3] The later government and foundation supported undertaking, encouraged by the Cold War and government concern for the cost of intensive language instruction, led to two major studies, one of which reached a much higher level of practical usefulness.

The send point is that, by the late thirties, it was widely and clearly understood that aptitude, however defined and however precisely measured, could only account for part of the variance in language learning success. The fuller instructional model set out by Carroll (1962), but understood in general terms at least thirty years earlier, show clearly that the various kinds of aptitude interacted with other personal factors (such as motivation) and with the instructional conditions to produce various kinds of success in language learning. In fact, by the 1930s, all of the items that might be included in this fuller model had been mentioned, so that the task was not to think of new ones, but to show the contribution of each to the model.

Prognosis testing

Our story then starts some sixty or more years ago. While egalitarian principles demanded that everyone should have the right of access to a high school education, including foreign language classes that were offered in them, the tiny amount of time allocated in the US school curriculum to language study led to a distressingly high failure rate. Motivated by what Michel (1936) referred to as 'the deplorable mortality in foreign language classes,' language testers set out to develop what they called prognosis tests,

[3] In fact, most if not all the ideas proposed at the Symposium as relevant to aptitude had been mentioned before 1942; what had not been done had been to show the exact weight to be given to each feature, but the Symposium papers did not do this either.

which, they hoped, could provide information about how well someone would perform in a language learning situation, or more precisely, about how to keep prospective failures out of their classes.

The genesis of prognosis tests was strictly practical rather than theoretical. Once it had become accepted in the USA in the early 1920s that general intelligence tests could be used with some effect to forecast how well a student would do at school, it was inevitable that some people would start to ask about the possibility of predicting success in specific subjects, including language study. This could then be used to alleviate the problems of teachers who felt themselves required to deal with students they believed unqualified for language study and who had been admitted to their classes through a policy of mass education.

This concern was highlighted in a paper entitled 'Mortality in modern languages students' by Cheydleur (1932a) reporting a long-term study of drop-outs and failures in language classes at the University of Wisconsin. After painting a picture of language departments agonizing over the numbers of their students who dropped out or failed their courses, Cheydleur argued for the value of using intelligence, placement and advancement tests to control student access and progress. Between 1925 and 1930, three prognosis tests for school use were prepared that stayed on the market for many years.

From the beginning, these tests combine two separate approaches to testing aptitude, which might be labeled the **analytical** and the **synthetic.** The analytical approach was to use items that tapped specific hypothesized cognitive abilities, usually through the first language, such as memory or vocabulary or some other aspect of verbal intelligence.

The synthetic approach was to give the candidate a mini-lesson in an artificial or foreign language, assuming that one could generalize from this short experience to performance in longer learning programs.

One of the earliest tests was written by Stoddard and Vander Beke, which included six subtests, three involving English grammatical skills -- singulars and plurals, tense,

nominalization, and three to do with guessing Esperanto words, applying Esperanto grammatical rules, and translating Esperanto sentences into English. A second was the *Language Aptitude Test* prepared by a team at George Washington University (Hunt et al., 1929), which involved learning elements of an artificial language. A third was the *Luria-Orleans Modern Language Prognosis Test* (Luria and Orleans, 1928), which took 85 minutes contained a language learning trial, consisting of vocabulary exercises (cognates and memorization) and eight grammar translation lessons in French and Spanish.

Prognosis in the Modern Foreign Language Study

It was while these early tests were being developed that the field of foreign language teaching was subjected to a major review by the Modern Foreign Language Study and the Canadian Committee on Modern Languages that started work in 1924 and went on for some years. The members of the committee were ardent supporters of prognosis:

> This Committee felt that no part of its experimental program would be more welcome to its colleagues as likely to throw light on their problems and bring relief from the difficult and often hopeless situation created by the numbers and unfitness of students, and it arranged, therefore, as soon as the foreign language achievement tests were well under way, to sponsor experimental undertakings in the field of prognosis. (Henmon, 1929:v).

The motivation was fundamentally economic, the goal being to replace 'wasteful methods of trial and error' with more efficient selection of students and their assignment 'to the work for which they are best fitted.' (Henmon 1929:3) The problems studied by the eight researchers whose work was supported and reported by the Committees turned out in the event, however, to be extremely resistant to solution and their studies were discouragingly inconclusive. In the long run, they failed to

> bring evidence that any test has yet been devised which can be counted on to reveal linguistic incapacity or to show itself as a reliable instrument for selecting

successful students of foreign languages. The question of language prognosis is far too complex for such a categorical answer. (Henmon 1929:vi)

The theoretical question underlying the design of a selection technique was whether the mind should be conceived of as a 'host of highly specialized capacities which may vary independently' or as 'a unitary affair' with the various parts correlated and forming 'a common factor of general intelligence.' American educational psychology, Henmon noted, was inclined to the belief in a high degree of specialization, which was why the search for specific abilities was being so enthusiastically pursued. He saw the task as being to determine the relative contributions of general intelligence and special aptitudes to predictions about student performance.

The belief in the importance of special aptitude was well entrenched in the profession. Two-thirds of the US and Canadian modern language teachers questioned in a 1926 survey had found cases of students with 'linguistic disability or incapacity not accompanied by low general intelligence.' Intelligence was believed to be a factor. Henmon saw it as the task of his research group to answer four basic questions:

- Is there a minimal IQ level for successful modern language study?
- Is there a minimal general scholarship level for successful modern language study?
- Can special language learning abilities be recognized, tested, and used for prediction of success?
- Can one semester's results be used to predict future success?

The work of the 1920s was reported in a book published by the Modern Language Study (Henmon *et al.* 1929). In the introductory essay, Henmon summarized recent work looking at correlations of intelligence quotients and scores with school marks or objective test scores in modern languages. Most of the studies had shown a low positive correlation,

ranging from 0.20 to 0.60, not much use for practical decision making. The 'variability, inaccuracy and subjectivity of school marks' were so well established that they could not be expected to help much. But Henmon was convinced of the value of the continued search for special language abilities.

In the first of six reports of current work, John Bohan looked at the relation between scores on intelligence tests given to entering students at the University of Minnesota between 1921 and 1925 and their later grades in English and Foreign Languages, finding correlations between 0.15 and 0.50.

Carl Brigham, teaching at Princeton University and already associated with the College Entrance Examination Board where he was developing the SAT, studied the Princeton artificial language test invented by Stuart Dodd, which had been shown to have high diagnostic validity as a general intelligence test but the prognostic adequacy of which was limited. Brigham analyzed various correlations in the case of 236 men for whom there were full enough data. The best predictor of college French marks was the average of College Board Entrance Examinations in French, English and Latin (0.480); neither the intelligence test (0.276) nor the language test (0.269) were nearly as useful predictors, nor did the latter two tests add much to the prediction of the examinations (0.533).

In another chapter, L. Thomas Hopkins, at the University of Colorado, found the Wilkins Prognosis Test and the Wilkins Elimination test to be 'a reliable measure of some kind of ability or particular type of function,' but not of the ability to succeed in foreign languages.

George Rice, at the University of California, gave a test written by May Barry which taught some Spanish grammar items and vocabulary to 100 pupils as a trial experience in language learning. The test correlated with intelligence quotient (0.79), and with teacher's marks at the end of the year (0.60) better than the intelligence score did (0.53).

Percival Symonds, teaching at Teachers College, Columbia University, whose test was later used in a number of studies and must have been widely accepted, pointed out the problem of determining the value of a prognostic test. Even if such a test could measure aptitude, it was judged by its correlation with achievement, which was the combined result of aptitude, 'and of the forces of instruction, including interest and interest of the learner, organization of the material, skill of the teacher, etc.' This model, set out formally by Carroll (1960, 1962) is often forgotten or overlooked by researchers venturing into the area of aptitude testing for the first time. None of the earlier researchers ever claimed that aptitude alone accounted for research; there is nothing novel in the claims (heard even at the 1994 Symposium) that other personal or instructional factors need to be taken into account.

But recognizing this complexity made the validation of a prognosis test doubly difficult: first, a test will have normally been used to exclude unsuitable students from the course and so from the validation study, and secondly, the aptitude test is known to measure and account for only part of the assumed causes of later variation.

In spite of this problem, Symonds believed that three types of aptitude tests made sense: measures of general intelligence, tests of ability in the student's native language, and 'quick-learning tests in the new language.' He gave pupils in four schools a set of intelligence tests compiled by E.L. Thorndike, four quick-learning tests (two by Dodd and two using Esperanto by Symonds himself) and the Iowa Placement Examination (Foreign Language Aptitude). Those pupils who lasted the semester then took the American Council Beta French and Spanish Tests. While various problems with the skewing of some of the tests meant that the regression weights could not be relied on, the correlations suggested that those tests which included elements of translation ability (grammatical knowledge in particular) were likely to be good predictors of success in the classes.

In the final chapter, John Todd, a psychologist at the University of California, included in a test items based on a psychological analysis of the language learning process: a general questionnaire, a test of immediate auditory memory span for isolated digits, and

tests of the extent of native vocabulary and range of information. A number of studies were carried out. IQ was found to correlate well with school marks in languages. IQ tests also correlated well with Todd's linguistic test. Todd was satisfied that he had not found evidence of a special language aptitude: 'Whatever our tests may have measured it plainly was not a linguistic "talent" or special aptitude. If linguistic special aptitude is a reality, some other distinct type of test must be invented for the purpose of measuring it.' (1929:161) Todd's negative findings must have had a temporarily dampening effect on what ultimately proved to be the most useful avenue of research, namely the testing of much more specific abilities.

With the publication of the collection of papers by Henmon *et al.* (1929), the place of prognosis as a central topic in language testing research had been established, and the general model within which a solution must be found had been delineated, but there had been no widely accepted answers to the questions that had been raised.

The Symonds tests of prognosis

Over the next decade, research on prognosis continued. Symonds continued his research with the aptitude test that he had designed (1930a), reporting in a study (1930b) a correlation of 0.71 between the prognosis test and a later achievement test.

The effectiveness of the Symonds' Foreign Language Prognosis Tests was examined in a number of studies over the next few years. Richardson (1933) administered them to 242 high school freshmen planning to take foreign languages finding a correlation of about 0.60 with first semester scores. Richardson did find the prognosis tests gave better predictions than intelligence tests with two cohorts of 120 high school students.

In research for an MA thesis at the University of Chicago, Lau (1933) administered the test to eighty pupils in three Michigan high schools on their first day of class, and found a 0.60 correlation with the American Council Alpha tests at the end of the semester. The weakest correlation was with vocabulary and the strongest with grammar.

An elaborate study using both the Symonds' and the Iowa foreign language aptitude tests was undertaken as a master's thesis at the University of Minnesota, Sister Virgil Michel (1934, 1936), a teacher at the St. Joseph's Academy. She acknowledged her inspiration to the statement by Symonds that 'prognostic testing is the romantic chapter in the history of educational measurement,' and agreed also with the platitude that failing students should have been guided into easier classes, but noted that educational prognosis was 'still in its infancy.'(1936:275) She administered the Symonds Foreign Language Aptitude Test to a group of high school students and the Iowa Foreign Language Test to a smaller group of beginning college German students at the college level) and to both, a newly devised German prognosis test that she had constructed including a memory test of short German sentences with their English translations, an analogies test of words that were cognate in German and English, and a series of German grammar rules and exercises. For the high school students, none of these tests gave useful correlations with the Columbia Research Bureau German Test or with teachers' marks at the end of the first semester. Multiple correlations combining the tests did not help much. She concluded that the Symonds test using Esperanto seemed to have not done as well with German as with French and Spanish For the university students, combinations of the Iowa test (which also used Esperanto) and the German prognosis tests did achieve correlations with the end of semester marks, but not much better than did the high school average. Her thesis concludes somewhat pessimistically:

> In general, the experiment corroborates the findings of the majority of investigators in foreign language prognosis in so far as the correlations are rather low, in so far as predicting success in any one subject is much more difficult than prognosis of success in all subjects in high school or university, and in so far as it points with increasing insistence to the need for further research to secure more efficient predictive measures than those that exist at present. (Cited from Coleman and King, 1938:435).

'Romantic' as the topic may have been, there were no signs of a happy ending yet, but the Symonds Test continued to produce useful results for French high school classes,

and Sister Virgil's recognition of the possible language specificity was an important advance

Kaulfers on prognosis

If the correlations cited so far seem low, an even more pessimistic picture emerged from the work of the California foreign language education researcher, Walter Kaulfers (1931), who found IQ scores or English marks to be better predictors than standardized foreign-language aptitude tests. Kaulfer's work on prognosis formed the basis of his Ph.D. dissertation written at Stanford University (1933). Reviewing over 650 correlations, published since 1901 by nearly fifty researchers, between foreign language achievement and nearly seventy other factors, he found large variability. The medians for the most common factors were prognosis tests (0.60), English ability (0.46), general language ability (0.44) and mental ability (0.35). His work left Kaulfers unconvinced that there was a special language aptitude, and he judged the prognosis tests to be nothing more than weighted intelligence tests. Because of the unstandardized conditions in junior high school Spanish classes, he saw little likelihood of getting predictive efficiency of much higher than twenty to thirty percent. As early as Kaulfer's dissertation, then, it was fully understood that the effectiveness of an aptitude test was dependent on the instructional situation.

Kaulfers continued to think about prognosis. In a paper published in 1939, he again expressed a fundamentally pessimistic view, and concluded that 'prognosis as a panacean solution to foreign-language problems is destined long to remain in the limbo of wishful thinking.' The fundamental problem as he saw it was the proliferation of approaches to teaching: 'it is inconceivable that any one test, however comprehensive, could predict achievement in a field in which such a variety of methods, materials, and objectives abound.'

In the same year, Kaulfers wrote reviews of the Symond's *Foreign Language Prognosis Test* (2:1340)[4] and the *Luria-Orleans Modern Language Test.* (2:1341) The

[4]References are to Buros (1975).

former he considered to be no more than 'a linguistically weighted intelligence test,' to lack any validity data, and to achieve too low a prediction correlation to warrant its use to reject a student. In any case, its usefulness would be limited to grammar-translation courses, and it would be too difficult for any student below eighth grade level. The second test also appeared designed to predict achievement in 'the traditional grammar-translation type of course of a decade or more ago.' He had found its validity to be low, not enough to have any advantage over more easily available measures like a twelve-minute test of English vocabulary.

Kaulfers had put his finger on a key issue: a prognosis test measured not so much a general (or even a general special) ability as a number of abilities that would be of benefit in various language learning situations. Insofar as a foreign language teaching approach was focused on the same skills that were being used in other subjects, a simple native language vocabulary test would be as good as anything else as a predictor. Aptitude, then, while a matter concerning the individual pupil, could only be defined in the context of the teaching method that was to be used.

Other pre-war studies of prognosis

The study of language aptitude and of the possibility of predicting achievement in language learning continued to be a matter of considerable academic and professional interest for the decade after the publication of Henmon *et al.* (1929).[5] It was a popular topic for theses and articles, but there was no breakthrough. Many possible predictors were investigated, such as age, attitude and personality.

In Britain, there were some beginnings of interest in prognosis in Scottish Council for Research in Education Examination Inquiry (1934) that showed that, in French, university class marks were slightly better predictors (0.69) of degree marks than were

[5]The second volume of the *Analytical Bibliography* listed seventeen items dealing with prognosis, including Walter Kaulfer's doctoral dissertation discussed above, and the third volume, covering the years 1937-1942 but its publication delayed until after the war, (Coleman, King et al., 1949) listed twenty-five items.

secondary school teachers' estimates of the Leaving Certificate Examination administered by the school (0.55).

One paper that appeared in 1939 looked ahead to much of the work that was to come. Spoerl (1939) asked what in fact constituted language learning ability? Was it intelligence, or courage, or form-color preference, or memory? She tested thirty-eight advanced German students at the American International College in Springfield, Mass., on the Henmon-Nelson test of mental ability, the Allport Ascendance-Submission Reaction Study (to test attitude and openness to suggestions in the new foreign language situation), and the Revised Minnesota Paper Form Board Test (to see if form recognition was relevant), and had them also given the Co-operative German Test. Major differences emerged between men and women: the correlation between class grade and Cooperative test score was 0.35 for men and 0.73 for women; similarly, the correlation between the intelligence measure and the grade was 0.63 for women and 0.123 for men. Neither the test of forms nor the ascendance submission test had significance relation to the German scores. Her conclusion was that while intelligence was significant for women, it was not for men.

Looking back over the first decade's work in prognosis testing, evidently the earlier expectation of Henmon and the Modern Foreign Language Study had not been met. An article by Tallent (1938) was recorded by Kaulfers as the sixtieth article published since 1901 showing that 'prognostic testing cannot be depended upon to solve foreign language problems.' Prophesy, it seemed, was dead.

A more dispassionate reconsideration suggests that the researchers of the period had helped clarify the issue enormously, and recognized the limitations of their task in that they were being asked to predict a more or less immeasurable attainment in uncontrolled and variegated learning situations. They were aware of the problems caused by the variation in goals and methods of teaching contexts, cognizant of the need for multiple rather than single predictors, and open to the complexity resulting from the fact that aptitude (however measured) was only one of a number of factors accounting for

achievement. The results of tests that they had developed, which were either slightly modified intelligence tests or mini-lessons in language, when used together with other available data, did permit a wise high school counselor to give useful advice to students identified as unlikely to succeed in formal language learning classes, and did permit responsible schools to make special provision for pupils who would be unlikely to benefit from such classes. Their tests were, as Carroll (1960) concluded when he started his own major work, 'reasonably effective in predicting success' in classes whose main objective was teaching the ability to read and translate a foreign language. They were to prove much less effective in predicting performance on more communicatively oriented programs, a challenge that was to be met by Carroll and others a quarter of a century later. But given the limited support for the research they had tackled, the high level of understanding reached during this first period deserves better recognition.

The Army UCLA aptitude study

The issue of prognosis did not die. During the war, admission to intensive language training courses in the military forces was based mainly on previous education. Frith (1953) reported at the 1953 Georgetown Round Table that the Air Force used scores on general intelligence and technical aptitude tests, possession of a high school diploma and a desire to study the language as the criteria for starting the study of Mandarin Chinese.

With the peace-time need for more economically sound approaches, the issue of which people to train became significant. Frith (1953) described trial courses conducted as screening devices at the Air Force Institute of Technology. Morgan (1953) reported that another government agency used the same approach, but Morgan himself believed and claimed to have demonstrated that an hour's careful study by a clinical psychologist of material collected with a battery of tests, including a projective "written interview questionnaire" and a personality inventory, would produce equally valid predictions.

As language training developed in the post-war years at the Army Language Training School in the Presidio of Monterey, the possibility of saving wasted time and effort persuaded the Army to fund the construction and validation of foreign language aptitude tests. The contract for the study went to three psychologists at the University of California, Los Angeles. The project, led by Roy M. Dorcus assisted by George E. Mount and Margaret H. Jones (1953) lasted from June 1950 to May 1953 and dealt with six languages, Russian, Hungarian, Serbo-Croatian, Arabic, Japanese and Mandarin Chinese.

A preliminary search of the literature produced 'no studies of value in the design of language aptitude tests for the selection of language trainees,' apart from some results of the language portions of the West Point Qualifying Examination. The report did not discuss any of the large body of pre-war work on foreign language prognosis described earlier in this chapter and it is not clear whether the authors knew of its existence and considered it irrelevant, or whether as psychologists coming to the field from outside they were unaware of foreign language testing literature that could have given them a jump start in their work. Analysis of data routinely collected at the Army Language School revealed that only pitch correlated significantly with any of the language proficiency scores, and that only for the first written and the first course oral examinations.

Nonetheless, encouraged by the high correlation between early and late language scores to believe that there must be measurable aptitude facts that could help predict later results, the team developed a list of ten 'major aptitude skills' which could be measured with a group pencil-and-paper test; this latter limitation prevented the testing of oral manipulation skills. The items chosen show a psychologist's rather than a linguist's view of the process of language learning. Perhaps if Harvard had been closer to Monterey, a more qualified research team might have been selected -- it was on the grounds of distance that John Carroll's bid for the contract was turned down. (Carroll, personal communication, 19 October 1993)

The test battery, different for each language, was administered to 150 incoming trainees in 1950 and scored at the University of California, Los Angeles, and compared

with proficiency scores on a complete battery of language proficiency tests also constructed by Army Language school staff for the study. The results of the study were disappointing. The West Point Qualifying examination continued to be the best predictor of the outcome of training, about 5-10 per cent above chance. Adding the selection tests did not improve the predictive power much. While there continued to be evidence of aptitude in the high correlation of early and late scores, the various aspects measured appeared 'to include a relatively small part of the aptitude and skill required in the learning of a language.' While still convinced of the existence of language aptitude, the researchers had failed to find a way to measure it.

This was surely not the first, not will it be the last time that experts from a related field have failed because of their lack of understanding of language and their unwillingness to start from the current state of knowledge in the field of language learning. Unhampered by knowledge of earlier work, they were able to repeat mistakes and look in the wrong places.

The prediction of success in intensive foreign language training

A much more systematic attack on the problem of language aptitude was made by John Carroll, in some years of research funded by the Carnegie Foundation and conducted at the Laboratory for Research in Instruction, Graduate School of Education, Harvard University. Carroll reiterated the economic basis for the concern, because of the expense of the intensive language programs that required eight to twelve months of full-time study and which were being offered in programs like the Army Language School at the Presidio of Monterey. An accurate measurement of foreign language learning aptitude should be able to provide a valuable screening device for costly governmental programs and minimize training failures, which ran as high as 80 per cent in one Japanese program that had been studied by Williams and Leavitt (1947).

Carroll premised his investigation on two 'propositions.' The first was that the facility to learn to speak a foreign language is 'a fairly specialized talent (or group of

talents)' independent of the traits included under 'intelligence.' The second was that it is rare enough in the general population to make it worthwhile to be selective in choosing people for expensive intensive programs. Intelligence tests, he pointed out, had been relatively unsuccessful in screening people for language training. Even with groups carefully selected for general intelligence, Frith (1953) had found that trial courses led to the rejection of as many as 75 per cent. of the students The prognosis tests tried in the 1920s and 1930s had generally been limited, Carroll noted, to pencil-and-paper testing of English language ability or work-sampling of short lessons in cognitive, intellectual aspects of formal language learning. These tests, which generally correlated quite highly with intelligence tests, were often reasonable predictors of learning to read and translate but they had less relevance to learning to speak a language in an intensive course. Dorcus and colleagues, Carroll graciously suggested, had 'just missed' measuring the crucial abilities, in that their tests failed to tap the relevant abilities. Memory for digits, for instance, which they tested, was not relevant to language learning, while memory for sound, which they did not test, probably was significant.

Carroll started with an initial battery that contained twenty separate tests, each intended to check one of five factors of verbal ability that had been proposed by French (1951): verbal knowledge, word fluency (knowledge of orthographic habits), fluency of expression, associative memory, and naming. Also included was a Phonetic Discrimination task developed by Stanley Sapon that asked the subject to identify the odd sound out in a triad.

Carroll tried several kinds of work-sample tests. One was an artificial language test in which subjects learned the names of a simple foreign language number system. Another was a tape recording with accompanying film strip that taught a simple artificial language. A third presented a more formal artificial language through grammar lessons.

In this approach, Carroll was working on the same double strategy followed by earlier aptitude testers. If he could, he wanted to find tests that tapped the most basic

abilities in language learning, the discrete primary skills. Failing this, he sought to find the smallest trial learning situation that would predict performance in a full course.

The new tests were tried in a number of situations. In February 1954, 111 men pre-screened for admission to an eight month intensive course offered for the U.S. Air Force at Yale University took a four hour battery of tests. They then went into a three day preliminary training period, during or after which thirty-one withdrew voluntarily. The validity analysis was based on the remaining eighty, only thirty-three of whom were selected for the full course. Using as the criterion measure either grades given by instructors or the selection decision, a large number of test variables showed significant correlations. The summed results of four tests (artificial language learning, phonetic association, words in sentences, and paired associates) produced a multiple R of 0.74. The prediction test and the trial course had agreed in sixty-six out of eighty cases.

A second trial was carried out in June 1954, using some new types of items. Once again, validity coefficients were remarkably high, a multiple R of 0.77 -- and, using some of the new tests, 0.839. On the basis of these successes, the *Psi-Lambda*[6] *Foreign Language Aptitude Battery* was made available to the Air Force in 1955 for further testing, with generally satisfactory results. The screening policy finally adopted by the Air Force was to use the result of the battery as a criterion for admission to the trial course, and make a further cut after that.

Two series of tests were conducted to check the relevance of the battery for different types of languages. While the correlation in one sample was lowest in predicting success in learning languages with characters (Japanese, Chinese, Korean), this did not show up in a second sample. This result and other analyses supported the hypothesis of the non-specificity of language aptitude. The battery seemed to predict oral and written skills equally well, depending on the instructional approach.

---

[6]An abbreviation, Carroll noted, for psycholinguistic.

Experimental testing was also conducted at the Foreign Service Institute of the U.S. Department of State. Good correlations (about 0.70) were found with instructor grades in six-month long courses in twelve different languages. In another test, eighty-three trainees at the Foreign Service Institute were given the battery, which achieved a multiple R of 0.778 with performance at the end of a six month course. The test was much better than the prediction based on a fifteen-minute 'diagnostic interview' given to the candidates by the chair of the language department in which he was to study. The results of this study also produced evidence of the effect of age; while the subjects' aged showed a slightly negative linear correlation with their success in language learning, the fact that adding the age variable to the aptitude test did not improve the prediction showed that the aptitude test measured whatever in the age variable was relevant to success in language learning; it further contradicted the notion that older people cannot learn foreign languages successfully.

Carroll (1960) reported two situations in which the aptitude battery failed to make significant predictions. Sixty two persons in six month courses conducted by the National Security Agency were given a battery of tests before they began courses (typically six months long); the tests failed to predict their grades in these courses, which were concerned with the use of foreign language skills in "cryptanalysis and related matters." Carroll explained this as a result of the criterion being "poorly defined" or "irrelevant."(It is likely that Carroll was given no further details of the course or of the criterion tests. The National Security Agency tended to be security-conscious; as I recall, its linguists used to pretend to be working for the CIA.) In the second case of failure that he reported, the battery was given to two classes of U.S. Air Force personnel learning Russian in an intensive program in a charitably unnamed American university. Carroll attributed the lack of correlation between the battery and the criterion grades to the inconsistency of the latter scores, as well as to such associated matters as "the quality of the teaching, the quality of the text materials, and the reliability of the grading." From all these studies,

Carroll was satisfied that he had good evidence that the tests in the battery were "generally speaking, highly valid."

The Modern Language Aptitude Test

Given the general success of the battery, a commercial form of the Carroll and Sapon test was published in 1959 by the Psychological Corporation under the name, *Modern Language Aptitude Test.* In this form, it was tried out in the summers of 1958 and 1959 with students in intensive eight week summer courses in Arabic, Persian, Turkish or Modern Hebrew, producing correlations of about 0.5 with final grades.

In a major paper reviewing his work in developing successful aptitude measures, Carroll (1960) raised a more fundamental question. His studies to date had assumed that success was a direct function of measured aptitude. Such a model was 'oversimplified, if not downright wrong.' A better model would take into account other relevant factors, such as motivation and instructional variables. He proposed a model that included at least two instructional variables (adequacy of presentation and the time allowed for learning) and three individual variables (verbal intelligence, aptitude -- or amount of time needed to learn -- and motivation -- or the amount of time the learner would apply himself to the task. Using the resulting model, Carroll was able to demonstrate how variation in the conditions of the various courses accounted for variation in the predictive ability of the aptitude battery. Because aptitude is not the only variable accounting for success in language learning, its validity can only be shown when the other factors are taken into account.

In summing up his major study, Carroll concluded that language aptitude consisted of the four distinct and measurable abilities: phonetic coding[7] -- the ability to code an auditory phonetic signal so that it could be remembered for more than a few seconds, grammar handling[8]-- the ability to recognize functions of words in sentences, rote

[7]The Phonetic Coding Factor, Carroll (1993:171) notes, may be identical to the Spelling Cluster of abilities.

[8]It is still not clear, Carroll (1993:176) remarks, if the Grammatical Sensitivity factors represent a learned ability.

memorization ability of a large number of foreign language items,[9] and inductive language learning ability.[10] With the completion of this major body of research, then, Carroll could feel reasonably confident that he had managed to identify and measure the chief factors involved in aptitude for learning to speak a foreign language. His tests were able to account for most of the variation that could reasonably be attributed to aptitude.

While Carroll and Sapon's work did include validation of the use of the test in high school situations, the main goal of their test was to predict success in intensive courses of the kind more likely to be used at university level or for adults. A number of years later, Paul Pimsleur translated his findings into a published test battery, The *Pimsleur Language Aptitude Battery*.

The state of prophecy

When the Temple was destroyed, the Talmud says, the power to predict the future was taken away from prophets and given to fools and children.[11] Henmon and his colleagues' initial hope of achieving close to perfect prognosis was, it is now clear, over-optimistic. But they managed to show, and Pimsleur confirmed, that verbal intelligence tests do a good job in predicting not just how well a student will do at school, but how well he or she will do in typical foreign language classes, making it possible to schools to exclude students who are probably going to fail.

John Carroll added three vitally important dimensions. First, more successfully than anyone, he developed tests that measured, as well as anything can, some of the components of individual variation in ability to learn to speak a foreign language. The items in the *Modern Language Aptitude Test* continue to show up as robust factors in

---

[9]The memory factors identified in the aptitude studies appear to be special. See Carroll (1993:297-298).

[10]A more general foreign language ability factor may emerge, Carroll (1993:176-7) now says, if the test battery does not permit the Grammatical Sensitivity and the Phonetic Coding factors to emerge.

[11]Babylonian Talmud, Tractate *Baba Bathra*, 12b

studies of second language learning.[12] Second, he proposed a model that showed how measurable abilities interact with goals and methods. Third, his extended model made the whole issue clearer, by showing that aptitude was only one of the factors involved in what I have called a general theory of second language learning (Spolsky 1989).

Ultimately, then, the work on prognosis in the 1920s and 1930s and on language aptitude in the 1950s produced tests that could be used cautiously for selecting promising language students, and it provided, perhaps more important, an improved understanding of the nature of second language learning. Aptitude, this work clearly showed, is only one of the factors that can be used to predict success in second language learning. In seeking to make further advances in the field, it is unwise not to build on the work of our predecessors.

References

Buros, O. K. (ed.) 1975. *Foreign language tests and reviews*. Highland Park, New Jersey, The Gryphon Press.

Carroll, J. B. 1960. The prediction of success in intensive foreign language training (final revision). Laboratory for Research in Instruction, Graduate School of Education, Harvard University.

Carroll, J. B. 1962. 'The prediction of success in intensive foreign language training' in R. Glaser (ed.): *Training research and education*. Pittsburgh, The University of Pittsburgh Press. 87-136.

Cheydleur, F. D. 1932a. 'Mortality of modern languages students: its causes and prevention.' *Modern Language Journal* **17**(2): 104-136.

[12]For example, in a study of language gains by 658 American students in four month study-abroad programs in Russia, Ginsberg (1992) found two MLAT tests show up as significant predictors for gains in listening and reading. Carroll (1993) provides a reanalysis of early studies. The *Modern Language Aptitude Test* is still, at this writing, in print and use.

Coleman, A. and C. B. King (ed.) 1938. *An analytical bibliography of modern language teaching, vol. II, 1932-1937*. Chicago, University of Chicago Press.

Dorcus, R. M., G. E. Mount, and M. H. Jones. 1953 (mistakenly dated 1952). Construction and validation of foreign language aptitude tests. University of California, Los Angeles, for the Adjutant General's Office.

Edgeworth, F. Y. 1888. 'The statistics of examinations.' *Journal of the Royal Statistical Society* **51**: 599-635.

Foucault, M. 1975. *Surveiller et punir: naissance de la prison*. Paris, Gallimard.

Foucault, M. 1979. *Discipline and punish: the birth of the prison*. New York, Vintage.

French, J. W. 1951. *The description of aptitude and achievement tests in terms of rotated factors*. Chicago, University of Chicago Press.

Frith, J. R. 1953. 'Selection for language training by a trial course' in A. A. Hill (ed.): *Report of the fourth annual roundtable meeting on languages and linguistics*. Washington, DC, Institute of Languages and Linguistics, Georgetown University. 10-15.

Henmon, V. A. C. 1929. *Achievement tests in the modern foreign languages, prepared for the Modern foreign language study and the Canadian committee on modern languages*. New York, The MacMillan company.

Henmon, V. A. C., J. E. Bohan, C.C. Brigham, L.T. Hopkins, G.A. Rice, P.M. Symonds, J.W. Todd, and R.J. Van Tassel (ed.) 1929. *Prognosis tests in the modern foreign languages: Reports prepared for the Modern Foreign Language Study and the Canadian Committee on Modern Languages*. Publications of the American and Canadian Committees on Modern Languages. New York, The MacMillan Company.

Hunt, T., F. C. Wallace, S. Doran, K. C. Buynitzky, and R. E. Scharz. 1929. *Language Aptitude Test: George Washington University*. Washington, DC, Center for Psychological Service, George Washington University.

Kaulfers, W. V. 1931. 'Present state of prognosis in foreign languages.' *School and Society* **39**(8): 585-596.

Kaulfers, W. V. 1933a. Forecasting efficiency of current bases for prognosis. Unpublished doctor's dissertation, Stanford University.

Kaulfers, W. V. 1939. 'Prognosis and its alternatives in relation to the guidance of students.' *German Quarterly* **12**(3): 81-84.

Latham, H. 1877. *On the action of examinations considered as a means of selection*. Cambridge, Deighton, Bell and Company.

Lau, L. M. 1933. The use of the Symonds' Foreign Language Tests in Beginning French. Unpublished master's thesis. University of Chicago.

Luria, M. A. and J. S. Orleans 1928. *Luria-Orleans Modern Language Prognosis Test.* Yonkers, N. Y., World Book Company.

Michel, S. V. 1934. Prognosis in the modern foreign languages. Unpublished master's thesis, University of Minnesota.

Michel, S. V. 1936. 'Prognosis in German.' *Modern Language Journal* **20**(5): 275-287.

Morgan, W. J. 1953. 'A clinical approach to foreign language achievement' in A. A. Hill (ed.): *Report of the fourth annual roundtable meeting on languages and linguistics*. Washington, DC, Institute of Languages and Linguistics, Georgetown University. 15-21.

Richardson, H. D. 1933. 'Discovering aptitude for the foreign languages.' *Modern Language Journal* **18**(3): 160-170.

Scottish Council for Research in Education Examination Inquiry 1934. *The prognostic value of university entrance examinations in Scotland*. London, University of London Press, Ltd.

Spoerl, D. T. 1939. 'A study of some of the possible factors involved in foreign language learning.' *Modern Language Journal* **23**: 428-431.

Spolsky, B. 1989a. *Conditions for second language learning: introduction to a general theory.* Oxford, Oxford University Press.

Spolsky, B. 1995. *Measured Words.* Oxford, Oxford University Press.

Stoddard, G. D. and G. E. Vander Beke 1925. *Iowa Placement Examinations: Foreign Language Aptitude.* Iowa City, State University of Iowa.

Symonds, P. M. 1930a. Foreign Language prognosis test. New York, Teachers College, Columbia University.

Symonds, P. M. 1930b. 'A foreign language prognosis test.' *Teachers College Record* **31**: 540-546.

Tallent, E. R. E. 1938. 'Three coefficients of correlation that concern modern foreign languages.' *Modern Language Journal* **22**(8): 591-594.

Williams, S. B. and H. J. Leavitt 1947. 'Prediction of success in learning Japanese.' *Journal of Applied Psychology* **31**: 164-168.

STYLES OF THINKING AND LEARNING[1]

Robert J. Sternberg
Yale University

Introduction

Why do so many people who fail in school succeed in life, and vice versa? Why do some people turn to law, others to medicine, and still others to accounting? And why do some of those doctors who were straight–A students in medical school fail their patients? Why is it that some gifted kids get straight A's in school, whereas others with equal abilities flunk out? And why do some people learn foreign languages easily, and others only with great difficulty? These are just some of the questions that can be addressed through an understanding of styles of thinking and learning.

What happens in life depends not just on *how well* we think, but also on *how* we think. People think in different ways, and moreover, our research shows that they overestimate the extent to which others think the way they do. As a result, misunderstandings can develop—among spouses, parents and children, teachers and students, and bosses and employees. Understanding styles of thinking and learning can help people prevent these misunderstandings, and actually come to a better understanding of each other, and of themselves.

What are Styles of Thinking and Learning?

A style is a way of thinking. It is not an ability, but rather how we use the abilities we have. We do not have *a* style, but rather a *profile* of styles. People may be practically identical in their abilities, and yet have very different styles. Consider, for example, three friends: Alex, Bill, and Curt (who are real people--only the names are changed).

Alex was a model student right through his senior year of college. He received outstanding grades, and went to a highly prestigious college. The first time his academic career faltered was when he was in his senior year of college. For the first time, he had really to think for himself. Up to then, he had been able to get A's pretty much by doing what his teachers told him to do. But his senior essay was an independent project, and now he found himself at a loss. He was fine so long as other people told him what to do, but he was in trouble when he had to come up with his own ideas. He probably could have if he really wanted to; he just didn't like doing it, and didn't

---

[1]*Research for this article was supported under the Javits Act Program (Grant #R206R00001) by the Office of Educational Research and Improvement of the US Department of Education. Grantees undertaking such projects are encouraged to express freely their professional judgments. This article, therefore, does not necessarily represent the positions or policies of the Government and no official endorsement should be inferred.*
This paper was presented as a plenary paper at the 1994 Language Aptitude Invitational Symposium sponsored by the Center for the Advancement of Language Learning, held at Rosslyn, VA, from September 25-27, 1994. It has also appeared in Language Testing, *12 (3) 265-291.*

feel comfortable departing from the path of others. So he was smart and able, just so long as there was someone to guide him.

Alex had thought about being an historian or possibly a writer. He certainly had the ability to follow either of these careers. But his style of thinking was much better suited to the career he actually chose; today, he is a contracts lawyer, and a highly successful one. When asked what he does, he describes his work as directed by others. Investment bankers decide on a deal, and then instruct Alex to draw up a contract. Thus, the bankers set the structure, and Alex works within it. But if the bankers decide to modify their deal, then they have to pay Alex to do it. So every time they have an idea or change one, they pay Alex. He has found a job that is a good fit for his style of thinking. The key thing to remember is that Alex had the ability to do lots of things, but found a career that was a good fit to the way he likes to use his abilities. As a result, he is happy with this career.

Alex is also happy in his personal life, which in many respects is compatible with his professional life. Alex and his wife have 2.5 kids (well, three actually), and live in a comfortable suburb in a major metropolitan area. They keep up with the Joneses, and take their cues in their life from what others do. They are happy to follow whatever the going trends are, and thus to take their direction from society at large. They don't much question why they do what they do, but rather fall into the patterns set for them by others.

Bill matched Alex in abilities, but not in school achievement. Bill's primary style is quite different from Alex's, and it is one that is less rewarded by the schools. Whereas most schools value an Alex–the bright kid who does what he or she is told–fewer value a Bill–a child who is bright but who wants to do things his own way. Indeed, children like Bill can end up being viewed as behavior problems, or as lacking in ability.

Bill's experience was the opposite of Alex's. He got a mediocre grade in his introductory science course. He came into his own when he was allowed to work independently and truly to come up with his own ideas. He first really began to feel successful when he started his career as a research scientist. As a scientist, he was in a position to come up with his own ideas–his own theories, his own experiments. He no longer had to follow the dictates of a teacher or of anyone closely supervising what he did and how he did it.

Bill's personal life has also reflected his style of thinking. Bill's first marriage, to the "right" woman from the "right" background, ended in divorce. The marriage became the perfect image of what society says a marriage should be, but Bill was bored out of his mind. He had the right house in the right neighborhood with the right schools, and his wife thought he was crazy to be dissatisfied. In his second marriage, however, Bill leads the kind of lifestyle that he himself prefers, in the wrong neighborhood with the wrong spouse, and he is happy as he has never been before. He's doing it his way, which is what he always wanted.

Curt was similar in abilities to Alex and Bill. But his predominant style was different. As a college student, he was editor of the college course critique, and thus was in charge of evaluating every course taught at the college. When he went out on dates, he even gave his dates a test of

values—which they did not know they were taking. If they passed the test, he continued to go out with them; if not, that was the end of that relationship. Today, Curt is in his mid-40s, and perhaps predictably, is still not married.

But Curt, like Alex and Bill, has found a job that is a good fit to his predominant style of thinking. Curt always liked to evaluate people and things, and today he is a highly successful psychotherapist who evaluates people and their problems, and prescribes courses of therapy for them. Curt had the ability to do many things, but he found a job that was a good fit to his style of thought.

Alex, Bill, and Curt are the lucky ones. But go to any high school or college reunion, and you will meet scores of people who went into the wrong job for themselves. They may have done what their guidance or career counselor told them to do, based on abilities or even interests, but many of them have found careers where they feel like they are at a dead end. Being at a dead end is often in the mind of the beholder, and one often feels at a dead-end when the work one does is a misfit to the way in which one best uses the talents one has. Understanding styles can help people better understand why some activities fit them and others don't, and even why some people fit them, and others don't.

The Nature of Styles

Before learning about the styles themselves, one needs to know about some of the basic characteristics of styles. Consider a few of these basic characteristics.

First, as mentioned above, styles are not abilities but rather ways of using abilities. For example, two people could be equally smart, but one could prefer to work by him or herself, another to work with others. One could prefer to concentrate on the forest and ignore the trees, another to focus on the trees and not pay so much attention to the forest; or one might prefer to do things in novel and unconventional ways, whereas another might prefer the tried and true. In each case, people exploit the abilities they have in different ways, and often for different ends.

Secondly, people, including both parents and teachers, often confuse styles with abilities. For example, the teacher may view the child who has trouble following directions as stupid; or the teacher may see a child who is highly critical of the school as rebellious. The parent may view the child who is often off on cloud nine as unable to focus and concentrate. We need to understand styles so that we do not unfairly penalize bright people who just happen to have a style that is different from our own, or from one we value.

Thirdly, styles can vary from one task or situation to another. Although people have preferences in styles, they can't always follow these preferences. When you do your income taxes, you have to be very detail–oriented, whether you like it or not. When you work in a group, it pays to be attuned to other people, even if your normal tendency is to be a loner. Indeed, people need to learn how to be flexible, despite their preferences. In other words, the people who are most successful are usually those who can modify their style to fit the situation at hand. They are not rigidly bound to any one style, but rather flexibly adapt as the situation requires.

Fourthly, styles are socialized. In other words, we are not born with a fixed set of styles, to which we are doomed to adhere for the rest of our lives. Rather, we acquire styles by modeling those around us, such as parents, teachers, other authority figures, and peers. It is for this reason that what we do is so much more important in the development of our children than is what we say. Children model how we act rather than how we say they should act. You can't tell a child to pay attention to the needs of others, and expect the child to become attentive if you are not attentive yourself.

Fifthly, styles can change over the course of one's lifetime. Some people become more conservative in their thinking, for example, and others more liberal. Some become more global and holistic, others more detail–oriented. People are not locked into any one style or set of styles. Rather, they change as life's circumstances and their own predilections change.

Finally, styles are not better or worse, but merely different. In thinking about styles, we need to free ourselves of the mode of thinking that is customary when we think about abilities. A higher level of one style or another is not, in and of itself, better, although it may be more adaptive in a certain situation. However, the very opposite style may be more adaptive in another situation. What is better is not one or another style, but the flexibility to modify one's style as the situation demands. Here we can see how different styles are from abilities. We don't usually think in terms of less of an ability being better in a given situation—but this may well be the case for a style. For example, a style leading to one's being highly critical and evaluative may be better suspended when one is first trying to come up with new ideas. In the idea-generation stage, being too critical can result in one's accepting no new ideas at all.

Some History of the Concept of Styles

The theory of styles presented here is not the first one. Theories of styles were originally formulated when psychologists recognized the need for a bridge between their theories of abilities, on the one hand, and their theories of personality, on the other. Different kinds of theories have been proposed to address the need for such a construct.

For example, psychologist Jerome Kagan and others recognized that in their school work and in their lives, some children tend to be more reflective, others more impulsive. The reflective child, on the average, is at an advantage in school and in life. Impulsive children are too quick to believe that they are 'done' with a task, too quick to say what is on their mind, and too quick to follow the first idea that comes to them.

Another psychologist, Herman Witkin, noted that people differ in terms of their independence of the perceptual field that surrounds them. For example, some people, in an airplane, find it very difficult to detect whether they are upright with respect to the ground unless they are actually looking at the ground. Others can detect deviations from the horizontal easily and without looking at the ground. Similarly, some people can look at a painting and recognize symbols hidden or embedded in the midst of other objects, whereas other people have trouble discerning when some feature is embedded in the midst of others. Witkin and others found that those who

are able to differentiate themselves from the perceptual field generally are more successful in a variety of kinds of tasks and life pursuits.

Other theorists have taken a different tack. For example, Myers, following the lead of Jung, distinguished among eight different styles, divided into four groups of two. These styles are assessed in a widely used test battery called the Myers-Briggs Type Inventory. According to Myers, some people rely primarily on sensing the world around them, whereas others rely primarily on intuition. The sensing person tends to trust sensory experience more, whereas the intuitive person tends to trust her or his own intuitions, whether or not they correspond to sensory experiences. At the same time, some people are more oriented toward thinking, others toward feeling. The thinking–based person tends to prefer to approach problems logically and rationally, the feeling–based person to approach them on the basis of emotion. Myers further distinguished between orientations toward extroversion and introversion. The extrovert prefers to relate to and be with others, whereas the introvert prefers to be on his or her own. Finally, Myers distinguished between those emphasizing judgment and those emphasizing perception. The former tend to prefer to "process" data and to come to their own conclusions about what the data mean; the latter prefer to go with what they perceive and not to rely as much on their interpretations of the data.

As a last example, Gregorc has distinguished between sequential and random thinkers, on the one hand, and concrete and abstract thinkers, on the other. The sequential thinker is a linear thinker—someone who likes to start at Step 1 and to end at Step N. The random thinker, on the other hand, eschews linear, ordered progressions, and likes to jump around in his or her thinking. This individual is likely to chafe at the restrictions that conventional schooling places upon him or her.

Theories such as these set the stage for the theory of mental self-government, which incorporates elements of some of these past theories, while introducing new elements of its own.

The Theory of Mental Self-Government

Why a Theory of Mental Self-Government?

The basic idea of the theory of mental self–government is that the forms of government we have in the world are not coincidental. Rather, they are external reflections of what goes on in people's minds. They represent alternative ways of organizing our thinking. Thus, the forms of government we see are mirrors of our minds.

There are a number of parallels between the organization of the individual and the organization of society. For one thing, just as society needs to govern itself, so do we need to govern ourselves. We need to decide on priorities, as does a government. We need to allocate our resources, just as does a government. We need to be responsive to changes in the world, as does a government. And just as there are obstacles to change in government, so are there obstacles to change within ourselves.

The various styles in the theory are presented below. In order to make the description of each style more concrete, the characterization of the style will be preceded by three statements from a thinking-styles inventory that we use to assess people's styles of thought (Sternberg and Wagner, 1992). Readers can thereby evaluate the extent to which each style is typical of their own way of thinking.

For each statement, you can rate yourself on a scale from 1 to 9, where 1 means that the statement does not characterize you at all, and 9 means that the statement characterizes you extremely well. Intermediate points can be used for intermediate ratings. You can then evaluate yourself on the style by summing the three ratings for each style and dividing by three. Roughly speaking, a score from 1–3 indicates a low level of a style, a score of 4–6 indicates an intermediate level, and a score of 7– 9 indicates a high level.

The Functions of Mental Self-Government

Roughly speaking, governments serve three functions: executive, legislative, and judicial. The executive branch carries out the policies and laws enacted by the legislative branch, and the judicial branch evaluates whether the laws are being carried out correctly and if there are violations of these laws. People also need to enact these functions.

The Legislative Style

1. When I work on a project, I like to plan what to do and how to do it.
2. I like tasks that allow me to do things my own way.
3. I like to pursue tasks or problems that have little structure.

These three items measure the legislative style. Legislative people like to come up with their own ways of doing things, and prefer to decide for themselves what they will do and how they will do it. Legislative people like to create their own rules, and prefer problems that are not prestructured or prefabricated. In the examples of the introductory section, Bill was a legislative stylist. Some of the preferred kinds of activities of a legislative stylist are writing creative papers, designing innovative projects, creating new business or educational systems, and inventing new things. Some of the kinds of occupations they prefer, all of which let them exercise their legislative bent, are creative writer, scientist, artist, sculptor, investment banker, policy-maker, and architect.

The legislative style is particularly conducive to creativity, because creative people need not only the ability to come up with new ideas, but also the desire to. Unfortunately, school environments do not often reward the legislative style. Indeed, even the training for occupations that require people to be creative often discourages the legislative style. Thus, a person might find him or herself in a science course, required to memorize facts, formulas, and charts. Yet, scientists virtually never have to memorize anything; if they don't remember something, they look it up on their bookshelf.

The author observed an excellent example of how a physics teacher could teach in a way that would encourage children to think legislatively. The teacher was doing a unit on mass. The

teacher brought the students out into the faculty parking lot, and showed the students his automobile. He also gave the students a few basic tools, such as a yardstick. The students' assignment: to compute the mass of the teacher's automobile, using only their ingenuity and the few tools that the teacher made available.

Creative writers also need a legislative style, but a legislative style is not often encouraged, and is often discouraged in literature classes, where the emphasis in the lower grades is likely to be on comprehension and, in the upper grades, on criticism and analysis.

The Executive Style

1. I like situations in which it is clear what role I must play or in what way I should participate.
2. I like to follow instructions when solving a problem.
3. I like projects that provide a series of steps to follow to arrive at a solution.

These items measure the executive style, which is characteristic of people, like Alex in the first section, who prefer to be told what to do and how to do it. Executive people like to follow rules and prefer problems that are prestructured or prefabricated. They like to fill in the gaps within existing structures, rather than to create the structures themselves. Some of the kinds of activities they are likely to prefer are solving given mathematical problems, applying rules to problems, giving talks or lessons based on other people's ideas, and enforcing rules. Some occupations that can be a good fit to executive thinkers are certain types of lawyer, a police officer on patrol, builder of other people's designs, a soldier, proselytizer of other people's systems, and an administrative assistant.

The executive style tends to be valued both in school and in business, because executive stylists do what they are told, and often do it cheerfully. They follow directions and orders, and evaluate themselves in the same way the system is likely to evaluate them, namely, in terms of how well they do what they are told. Thus, a gifted child with an executive style is likely to do well in school, whereas a gifted child with a legislative style is likely to be viewed as nonconforming and even rebellious.

I recently saw an example of how a high school literature class, which I had thought would emphasize analysis of literature, could be taught in a way that promoted an executive style. The children were reading the *Odyssey*, which is certainly one of the premier works of literature in Western civilization. The whole class I observed, however, consisted of the teacher's asking the students to identify sources of quotations, and to recount the events in the chapter they had read, in the order in which the events had taken place. The teacher's emphasis, in his own words, was on 'close reading.' Thus, the students were encouraged to read carefully, but not necessarily to understand what they were reading.

Peer-group pressure encourages children to adopt an executive style as well, but with respect to the norms of the peer group rather than of the school. Thus, pressure from many sources can lead students to adopt this style.

The Judicial style

1. I like to analyze people's behavior.
2. I like projects that allow me to express my opinions to others.
3. I like tasks that allow me to evaluate the work of others.

These items measure the judicial style, as shown by Curt in the example above. A judicial person likes to evaluate rules and procedures, and prefers problems in which one analyzes and evaluates existing things and ideas. The judicial stylist likes activities such as writing critiques, giving opinions on things, judging people and their work, and evaluating programs. Some of their preferred kinds of occupations are judge, critic, program evaluator, consultant, admissions officer, grant and contract monitor, and systems analyst.

My son once commented to me that he hated history, and when I asked him why he hated it, he answered it was because he didn't like memorizing dates. Although the work of an historian is in large part judicial—the analysis of historical events—many children get the idea that the work is largely executive—remembering dates of events. As in science, therefore, some of the most able students may decide to pursue some other field, even though their style of thinking may be well suited not to their preparation for the career, but for the actual career itself.

Problems of mismatching are not limited to the school. In many businesses, including schools, lower–level managers are sought who have a largely executive style. They do what they are told, and try to do it well. People with such a style are often then promoted into the higher levels of management. The problem is that, in the higher levels, a more legislative or judicial style becomes desirable. However, many of the people with a more legislative or judicial style may well have been derailed early in their management careers, so that they never get to the higher levels of management. The result can be a higher level of management that appears to be a victim of the Peter Principle, but that in fact has fallen victim to promoting people to higher positions whose styles were suited for lower but not for higher levels of responsibility. Small wonder, for example, that many school administrators are reluctant to accept change. They obtained the positions they have because they did what they were told to do, not because they liked to decide what to do in the first place.

Styles can be important in personal as well as professional relationships. Consider some examples, and how they may or may not work well together.

A natural pairing in a personal relationship is a legislative person with an executive person. The legislative person tends to be the one who decides what to do, whereas the executive person tends to be the one who makes sure that it gets done. You need both in a relationship—someone to make some decisions and someone to enact them. A potential problem can arise if the legislative person becomes bored with the executive person as it is the legislator who always seems to be the 'ideas' person, or if the executive person starts to resent the legislative person for always trying to decide what to do. Two legislative people can do well together and maintain interest in their relationship, if they can work out ways to resolve the almost inevitable conflicts that will arise from two people who both want to be the one to decide things. Two executive people together

will tend to be a conforming and 'typical couple'—one that looks to others to decide what they should do and how they should do it. They are likely to follow whatever the fads are, and their way of distinguishing themselves will be to follow these fads even better than the next couple. Two judicial people can also work together and enjoy evaluating other people and their foibles. The danger comes if they start turning their judicial tendencies toward each other rather than toward the outside. Thus, knowing your style as well as your partner's can help you understand better what the potential strengths and pitfalls of a relationship are.

The Forms of Mental Self-Government

The theory of mental self-government specifies four forms: monarchic, hierarchic, oligarchic, and anarchic. Each form results in a different way of approaching the world and the problems with which it confronts us.

The Monarchic Style

1. I prefer to finish one assignment before starting another.
2. I like to devote all my time and energy to one project, rather than dividing my time and attention among several projects.
3. I like to put in long hours of work on one thing without being distracted.

A person with a monarchic style is someone who is single minded, driven, and often believes that the means justify the ends. The individual tends to oversimplify problems, and not to let anything get in the way of his or her solving a problem. Monarchic people can be relatively inflexible and unself-aware, so eager are they to focus on bringing a task to a successful conclusion.

Monarchic bosses are difficult to work with, because they tend not to take human considerations into account. If a task is supposed to be done, it's supposed to be done, without excuses or extenuating circumstances. When you are married to a monarchic individual, you usually know it quickly. Often you see little of the person, and even if you do see the person, his or her mind may be elsewhere. If you, rather than, say, work, are the subject of the person's obsession, you may find yourself drowned by unwanted attention: it can be difficult to find room in which to breathe.

Monarchic children can present a problem in school, because they usually want to be doing something other than what they are doing, and are likely to be thinking about the other thing while they are supposed to be attending to the teacher. Sometimes, their interests are best served when a teacher (or parent) brings whatever they are monarchic about to bear on other things they are doing. For example, a child who has a strong interest in sports but is not a reader may become a reader if given sports novels to read (as I did with my son). A child who loves cooking but not math could be given math problems to do that involve recipes. In these ways, the child may become interested in things that previously were of no interest.

The Hierarchic Style

1. When undertaking some task, I like first to come up with a list of things the task will require me to do and then to assign an order of priority to the items in the list.
2. Whenever I engage in a task, it is clear to me in what order of priority various parts of it need to get done.
3. When writing, I tend to emphasize the major points and to de-emphasize the minor ones.

The hierarchic individual has a hierarchy of goals, and recognizes the need to set priorities, as all goals cannot always be fulfilled, or at least fulfilled equally well. This person tends to be more accepting of complexity than is the monarchic person, and recognizes the need to view problems from a number of angles so as to set priorities correctly.

I once had a student who, when she would meet with me, always would have a list of things she wished to discuss. The items on the list were ordered in terms of priorities, and she would cross off items as she finished bringing them up. One day, she came in with a list that looked different from her usual one. I asked her whether she had changed her list, and she explained to me that she had come to have so many lists that she was now carrying around a list of lists. She was, without doubt, a hierarchical stylist!

Hierarchic individuals tend to fit well into organizations because they recognize the need for priorities. However, if their priorities are different from those of the organization, problems may arise. Then they may find themselves organizing their work according to their own, but not their organization's, priorities. The company lawyer who wants to spend too much time on *pro bono* work, the university professor who wants to spend too much time on teaching, and the cook who wants each meal to be perfect but who takes forever in cooking the meals may soon find themselves unwelcome in their respective organizations.

The Oligarchic Style

1. When there are competing issues of importance to address in my work, I somehow try to address them all simultaneously.
2. I sometimes have trouble setting priorities for multiple things that I need to get done.
3. Usually when working on a project, I tend to view almost all aspects of it as equally important.

Individuals preferring the oligarchic style are like hierarchic people in their desire to do more than one thing within the same time frame. But unlike hierarchic people, they tend to be motivated by multiple and often competing goals of equal perceived importance. Often, these individuals feel tense and even helpless in the face of competing demands on their time and other resources. They are not sure what to do first, or how much time to allot to each of the tasks they need to complete.

Minor interventions can often make the difference between success and failure for the oligarchic individual. F or example, a secretary was failing at her job because she was unable to get important tasks done on time. An oligarchic individual, she was as likely to do unimportant tasks early on as important ones. Many students have this same kind of problem, putting off the more important homework assignments in favor of the ones that are of lesser importance. The

secretary was about to lose her job when her supervisor tried one last intervention. For each assignment he gave the secretary, he assigned a priority score on a three-point scale. In this way, the secretary would have a clear numerical index of how important each task was. With this simple intervention, her work went from being poor to excellent. Similarly, oligarchic students can often do much better in school if parents or teachers help them set priorities for what needs to be done when.

The Anarchic Style

1. When I have to start to do some task, I usually do not organize my thoughts in advance.
2. When thinking about an issue that interests me, I prefer to let my mind wander with the ideas in whatever way it likes.
3. When talking about issues that interest me, I like to say things just as they occur to me, rather than waiting until I have organized or censored my thoughts.

The anarchic individual probably shows up as the least successful of the various stylists on a variety of tasks and in a variety of situations. This individual seems to be motivated by a potpourri of needs and goals that are often difficult for the anarchic individual, as well as for others, to sort out. The individual takes what seems like a random approach to problems, and to be driven by a muddle of seemingly inexplicable forces. The person has trouble adapting to systems because of a tendency to eschew any system at all, and to fight back at whatever system the individual seems as confining him or her.

Although anarchic individuals tend to have trouble adapting to the worlds of school and work, they often have greater potential for creative contribution than do many of the people who find the anarchics so distasteful. Because anarchics tend to pick up a little from here, a little from there, they often put together diverse bits of information and ideas in a creative way. They are wide-ranging in the scope of things they will consider, and so may see solutions to problems that others will not see. The problem for the teacher, parent, or employer is to help the anarchic person harness this potential for creativity, and achieve the self-discipline and organization that are necessary for any kind of a creative contribution. If this harnessing effort works, then the anarchic person can succeed in domains where others may fail.

The Levels, Scopes, and Meanings of Mental Self-Government

The three aspects of style to be considered here complete the theory. Each is considered in turn.

The Global Style

1. I like to do projects in which I don't have to pay much attention to details.
2. In any written work I do, I like to emphasize the scope and context of my ideas, that is, the general picture.
3. Usually when I make a decision, I don't pay much attention to the details.

Global individuals prefer to deal with relatively large and abstract issues. They ignore or don't like details, and prefer to see the forest rather than the trees. Often, they lose sight of the trees that constitute the forest. As a result, they have to be careful not to become lost on cloud nine.

The Local Style

1. I like problems that require engagement with details.
2. In carrying out a task, I am not satisfied unless even the nitty-gritty details are given close attention.
3. When writing, I like to focus on one thing and to scrutinize it thoroughly.

Local stylists like concrete problems requiring working with details. They tend to be oriented toward the pragmatics of a situation, and are down to earth. The danger is that they may lose the forest for the trees.

Global and local people can work particularly well together, because each attends to an aspect of task completion that the other would rather forget. Two global people trying to complete a project may each want to deal with the big issues, leaving no one to attend to the details; two local people may find themselves without anyone to do the higher order initial planning needed to get the job done. It helps if neither individual is so extreme that he or she cannot understand and appreciate what the other has to offer. Extreme localists or globalists can get carried away, and start to lose sight either that the big issues exist, or that there are details that someone needs to attend to.

The Internal Style

1. I like to be alone when working on a problem.
2. I like to avoid situations in which I have to work in a group.
3. To learn about some topic, I would rather read a well–written book than participate in a group discussion.

The internalist is concerned with internal affairs—that is to say, this individual turns inward. Internal individuals tend to be introverted, task-oriented, aloof, and sometimes socially less aware. They like to work alone. Essentially, their preference is to apply their intelligence to things or ideas in isolation from other people.

An example of how teachers can confuse style with abilities is shown by the case of a kindergartner who was recommended by her teacher for retention. When asked why she had made this recommendation, the teacher pointed out that although the child's academic work was quite good, the child did not seem 'socially ready' for first grade. That is to say, the child preferred to be on her own rather than to interact with other children, which the teacher took as a lack of some kind of social intelligence. In fact, the child was simply an internal. She was promoted, and has done splendidly well both academically and in her social relations.

The External Style

1. Before I start on a project, I like discussing my ideas with some friends or peers.
2. I like to work with others rather than by myself.
3. I like talking to people about ideas that occur to me and listening to what they have to say.

External individuals tend to be extroverted, outgoing, and people–oriented. Often, they are socially sensitive and aware of what is going on with others. They like working with other people wherever possible.

Many of the questions that arise in education as to 'what is better?' stem from a fundamental misunderstanding of the interaction of styles with learning experience. For example, in recent years, there has been a strong push toward what is called 'cooperative learning,' which means children working together to learn in groups. The idea is supposed to be that children will learn better in small working groups than they will when they are left to their own devices.

From the standpoint of the theory of mental self–government, there is no one right answer to questions such as whether children learn better individually or in groups, and, indeed, this question, like so many others, is viewed as misformulated. External children will prefer working in groups and will probably learn better when learning with others. Internal children will probably prefer to work alone, and may become anxious in a group setting.

This is not to say that internals should never work in groups, or externals, alone. Obviously, each kind of individual needs to develop the flexibility to learn to work in a variety of situations. But the styles point of view implies that teachers, like students, need to be flexible in the way they approach the teaching-learning process. They need to provide children with both individual and group settings so that children can be comfortable some of the time, and challenged the rest of the time. Always providing the same working setting tends to benefit some students, but to penalize others.

The Liberal Style

1. I like to do things in new ways, even if I am not sure they are the best ways.
2. I like to avoid situations where I am expected to do things according to some established way.
3. I am comfortable with projects that allow me to try unconventional ways of doing things.

The liberal stylist likes to go beyond existing rules and procedures, to maximize change, and to seek situations that are somewhat ambiguous. The individual is not necessarily 'politically' liberal. A political conservative could have a liberal style in trying to implement, say, a Republican agenda in a new and all-encompassing way. Thrill-seekers tend to have a liberal style, as do people who, in general, quickly become bored.

The Conservative Style

1. I like to do things in ways that have been shown in the past to be correct.
2. When I am in charge of something, I like to make sure to follow the procedures that have been used before.
3. I like to participate in situations where I am expected to do things in a traditional way.

The conservative stylist likes to adhere to existing rules and procedures, to minimize change, to avoid ambiguous situations where possible, and to stick with familiar situations in work and professional life. This individual will do best in a structured and relatively predictable environment.

The Development of Thinking and Learning Styles

Where do styles come from? How do they evolve? Styles seem to be largely socialized. From early on, we perceive certain modes of interaction with others and with things in the environment to be more rewarded than others, and we probably gravitate toward those modes. At the same time, we have built-in predispositions that place constraints on how much and how well we are able to adopt these rewarded styles. To some extent, society structures tasks along lines that benefit one style or another in a given situation. We therefore need to learn when to be what if we wish to adapt.

Consider some of the variables that are likely to affect the development of thinking and learning styles.

A first variable is culture. Some cultures are likely to be more rewarding of certain styles than of others. For example, the North American emphasis on innovation and making the 'better mouse-trap' may lead to relatively greater rewards for the legislative and liberal styles, at least among adults. Many national heroes of one kind or another in the USA, such as Edison as inventor, Einstein as scientist, Jefferson as political theorist, Steven Jobs as entrepreneur, and Hemingway as author, are heroes by virtue of their legislative contributions.

Other societies, such as Japan, that traditionally more highly emphasize conformity and the following of traditions, may be more likely to lead to executive and conservative styles. Perhaps, then, it is not so surprising that so many Nobel Prizes have been awarded to Americans and so few to Japanese. At the same time, the Japanese have found a way of maximizing on their own profiles of styles. Although many innovations in technology and elsewhere have not originated in Japan, the Japanese have attended to the details and taken the patterns of the originators to produce better products than the originators did—hence, their success in so many technological markets.

A second variable is gender. Traditionally, a legislative style has been more acceptable in males than in females. Men were supposed to set the rules, women to follow them. This tradition is changing, but it would probably be fair to say that many of the disadvantages women have experienced in the sciences, in business, and elsewhere have stemmed from their being labeled as

stylistically inappropriate when, say, they have given rather than followed orders. I believe that even today, young girls are socialized into stylistic roles (e.g., the executive role of doing what they are told) in a way that is to their disadvantage if they later try to make it in a variety of life pursuits.

A third variable is age. Legislation is generally encouraged in the preschool young, who are rewarded for being creative in the relatively unstructured environments of the nursery school and kindergarten. But after very short order, these same individuals are expected to become executive in the classroom, doing what they are told when they learn reading, writing, and arithmetic. Beyond kindergarten, children are in a situation where, for the most part, the teacher decides what the children should do and the children are expected to do it. Then, we who are college professors complain that the students we get at the college level don't want to think for themselves, and want to be told what to do. Of course! We made them that way!

A fourth variable is parenting or teaching style. What the parent or teacher encourages and rewards is likely to be reflected in the style of the child. The parent or teacher sets of model that children then emulate. It matters much less what we say than what we do. Children become what they see, not what they are told they should be.

Parents and teachers will not necessarily foster in children the same pattern of styles that they have themselves. All of us have encountered, for example, teachers who may like to think for themselves, but who may not want their children to do the same. Many political leaders show this same unfortunate trait. They want to decide on the policies for others to follow. The higher the need for power, the more the individual in charge may seek to develop conforming behavior among those who are expected to follow.

A fifth variable is religion. Some religions are more encouraging of questioning and confrontation than are others. Nobel Prizes, for example, are extremely unevenly distributed among the world's religions, and if population figures are taken into account, then the distributions are particularly striking. For example, Jews are way over-represented and Catholics under–represented in the awarding of these prizes. This is not to say that there is a right and a wrong way to exercise religion. It is to say that the way it is exercised will probably have profound stylistic effects that go beyond religious beliefs to the ways the child and later the adult think about the world in general.

Assessment, Schooling, and Mental Self-Government

How Styles are Assessed

We have used several converging operations to measure thinking styles, which have been reported in a series of journal articles and book chapters (Sternberg, 1990, 1993, 1994; Sternberg and Grigorenko, 1993; Grigorenko and Sternberg, 1995), but never before together in a single volume.

A first measure is the kind previewed above in the descriptions of the styles. This measure is called the 'thinking styles inventory.' People are given statements like 'If I work on a project, I like to plan what to do and how to do it' (which measures the legislative style), and rate the extent to which the style characterizes them on a 1-9 scale. We have computed normative data for this measure so that people can assess where they stand in relation to others.

A second measure is the 'set of thinking styles tasks.' In this measure, styles are measured via performance rather than merely by people passively evaluating statements. Consider an example. Imagine that you are the mayor of a small northeastern city. You have a city budget this year of $1 million. Below is a list of problems currently facing your city. Each would cost $1 million thoroughly to solve. Your job is to decide how you will spend the $1 million available to improve your city. Whether you spend all the money to solve one problem or divide up the money partially to deal with more than one problem is up to you.

1. Drug problem.
2. Roads.
3. Landfill.
4. Shelters for the homeless.

We are not interested in people's values. Rather, we are interested in their system of priorities. Scoring is on the basis of how funds are allocated. People who allocate all funds to one project are classified as showing a monarchic tendency. Those who set priorities in their distribution of funds are scored as hierarchic. Those who distribute money equally across projects are classified as oligarchic. And those who show no system at all are classified as anarchic.

A third measure, one of 'thinking styles evaluated by others,' is done on the basis of another person rating a first one. For example, a teacher might rate a student; a student, a teacher; a supervisor, an employee; and so on. Statements to be rated are ones like 'he or she prefers to solve problems in his or her own way' (to measure the legislative style) and 'he or she likes to evaluate his or her opinions and those of others' (to measure the judicial style).

A fourth measure assesses teaching and supervision styles, which, as mentioned earlier, may differ from the person's own individual style. Typical items on this scale are 'I want my students (employees) to develop their own ways of solving problems' (judicial style) and 'I agree with people who call for more, harsher discipline, and a return to the "good old ways"' (conservative style).

By using a variety of kinds of assessments, we are able to cancel out the biases and errors of measurement inevitably associated with a single kind of measurement, and thus better to converge on a more informed assessment of a person's profile of thinking styles. In the same way, readers will be able to gauge themselves in a variety of ways, to see the kinds of styles they use in different tasks and situations.

Styles in the Classroom

Elena Grigorenko and I (Sternberg and Grigorenko, 1993) have conducted several studies investigating styles in the classroom. One of these studies focused on teachers, another on students, and the third on the interaction between teachers and students.

In a first study with 85 teachers (57 female, 28 male) in four schools of widely varying types (private and public, and socio-economically diverse), we found several interesting effects with respect to grade taught, age of teachers, subject area taught, and ideology.

Teachers are more legislative but less executive at the lower grades than at the upper grades. These findings might suggest either that more legislative individuals are attracted toward teaching at the lower grade levels, or that people teaching at the lower grade levels become more legislative (or that those teaching at the upper grade levels are more executive). Either way, the demands on teachers in the U.S. are consistent with this pattern of findings: teachers in the upper grades are forced to follow a more rigidly prescribed curriculum than are teachers in the lower grades. The results are also consistent with our hypothesis that as children grow older, they are more and more socialized into an executive style of thinking, and away from a legislative style.

We also found older teachers to be more executive, local, and conservative than were younger teachers. Again, there are two interpretations of these findings, either or both of which might be correct. One interpretation is that teachers become more executive, local, and conservative with age; the other interpretation is that the difference is due to cohort effects. In other words, people of the earlier generations tend to be more executive, local, and conservative than people of the later generations. Either way, this constellation of traits is associated with authoritarianism in thinking, so that it suggests that older people become more 'fixed in their ways,' and that this change affects the way they interact with the young.

Further, we found that science teachers tended to be more local than were teachers of the humanities, whereas the latter tended to be more liberal than the former. These results again are roughly consistent with our experience. With respect to science, the results unfortunately suggest that science teachers may concentrate substantially more on the local details of science than on the 'big picture' of scientific research. Ironically, those students who may best be able to see the big picture may be those who are least appreciated by their science teachers.

Finally, we did an analysis of the relation of school ideology to teachers' styles (Sternberg and Grigorenko, 1993). We had a rater who was not familiar with the individual teachers in each school rate each school for its profile of styles on the basis of catalogues, faculty and student handbooks, statements of goals and purposes, and curricula. We also evaluated teachers' styles, and then did contrasts looking at the match between teachers and schools. For six of seven planned contrasts, we found significant effects. In other words, teachers tended to match the stylistic ideology of their schools. Either teachers tend to gravitate toward schools that fit them ideologically, or else they tend to become like the place they are in, suggesting again the importance of socialization in the formation of styles, even at the adult level. The suggestion is that we need to beware of the environment we enter, because we become like that environment.

In a second study (Sternberg and Grigorenko, 1993) of 124 students between the ages of 12 and 16 distributed across 4 schools, we found some interesting demographic effects. Socioeconomic level related negatively to the judicial, local, conservative, and oligarchic styles. These results are consistent with the notion of greater authoritarianism in the styles of individuals of lower socio-economic class. We also found that later-born siblings tend to be more legislative than earlier-born siblings, consistent with the past finding that first-borns tend to be more accepting of societal dictates than are later-borns. Finally, we found a significant degree of match between students' and teachers' styles. Whereas for the teachers, similarity of styles to the profile of their schools could be interpreted in terms of choice of school, such an explanation is implausible in the case of students, who rarely get to choose their school. The results again suggest socialization of styles.

In a third study, (Sternberg and Grigorenko, 1993) we went back to one of the original questions that motivated the work: do students do better in classrooms where their styles match rather than mismatch the styles of their teachers? We assessed students' and teachers' styles, and found that, indeed, students performed better and were more positively evaluated by teachers when the students' styles matched rather than mismatched the styles of their teachers. In other words, the students performed better when they were more like their teachers stylistically, independent of actual level of achievement. Teachers also tended to overestimate the similarity of students' styles to their own, probably leading them even more to teach in a way that would work for students who are similar to them, but not for those who are different. Clearly, then, the best teachers will be those who are flexible in their teaching and who meet the needs of students with diverse styles of thinking and learning.

Improving Instruction and Assessment

For those who teach and assess students at any level, or for those who have children who are taught and assessed, the theory of mental self-government implies modes of rendering teaching more effective. The key principle is that in order for students maximally to benefit from instruction and assessment, at least some of each of instruction and assessment should match their styles of thinking. I would not advocate a perfect match all the time: students need to learn, as does everyone, that the world does not always provide people with a perfect match to their preferred ways of doing things. Flexibility is as important for students as for teachers. But if we want students to show what they truly can do, match of instruction and assessment to styles is key.

Table 1. Thinking Styles and Methods of Instruction

Method of Instruction	Style(s) Most Compatible with Method of Instruction
Lecture	Executive/Hierarchical
Thought-Based Questioning	Judicial/Legislative
Cooperative Learning	External
Problem Solving of Given Problems	Executive
Projects	Legislative
Small-Group Recitation	External/Executive
Small-Group Discussion	External/Judicial
Reading	Internal/Hierarchical
For Details	Local/Executive
For Main Ideas	Global/Executive
For Analysis	Judicial
Memorization	Executive/Local/Conservative

Table 1 shows various methods of instruction and the styles that are most compatible with each of these methods. The major point of the table is that different methods of instruction work best for different styles of thought. If a teacher wants to reach and truly interact with a student, he or she needs the flexibility to teach to different styles of thinking, which means varying teaching style to suit different styles of thought on the part of students.

Table 2 shows various methods of assessment and the styles with which they are most compatible. Note that different methods of assessment tend to benefit different styles of thought. For example, multiple-choice testing is very much oriented toward executive and local thinkers. Thus, the enormous use of multiple–choice testing in the United States, allegedly to measure achievement and abilities, actually benefits people with one set of styles at the expense of people with other sets of styles. Styles are confounded with what is supposed to be measured, whether it be abilities or achievements. Projects tend to be oriented more toward legislative and judicial thinkers as well as toward global ones.

Note also the importance not only of the method of assessment used, but of the way in which the method of assessment is scored. For example, an essay can be scored for recall, in which case it benefits executive students, or for analysis, in which case it benefits judicial students, or for creativity, in which case it benefits legislative students. It is not the essay, per se, but how it is evaluated, that determines who benefits.

Table 2. Thinking Styles and Forms of Assessment

Form of Assessment	Main Skills Tapped	Most Compatible Style(s)
Short Answer/	Memory	Executive/Local
Multiple Choice	Analysis	Judicial/Local
	Time Allocation	Hierarchical
	Working by Self	Internal
Essay	Memory	Executive/Local
	Macroanalysis	Judicial/Global
	Microanalysis	Judicial/Local
	Creativity	Legislative
	Organization	Hierarchical
	Time Allocation	Hierarchical
	Acceptance of Teacher Viewpoint	Conservative
	Working by Self	Internal
Project/Portfolio	Analysis	Judicial
	Creativity	Legislative
	Teamwork	External
	Working by Self	Internal
	Organization	Hierarchical
	High Commitment	Monarchic
Interview	Social Ease	External

Finally, Table 3 shows how different prompts in instructional and evaluational assignments can lead to varying levels of compatibility for different styles. Prompts such as 'Who said...' and 'Who did...?' tend to benefit executive students; prompts such as 'Compare and contrast' and 'Analyze...' tend to benefit judicial students; and prompts such as 'Create. . .' and 'Invent...' benefit legislative students. By varying the kinds of prompts they use, teachers can equalize the benefits to all of the students whom they teach.

Table 3 Thinking Styles and Instructional/Evaluational Assignments (Style emphasized and types of prompts)

Executive	Judicial	Legislative
Who said?	Compare and contrast...	Create...
Summarize...	Analyze...	Invent...
Who did?	Evaluate...	If you...
When did?	In your judgment...	Imagine...
What did?	Why did?	Design...
How did?	What caused?	How would
Repeat back...	What is assumed by?	Suppose...
Describe...	Critique...	Ideally?

Styles and Second-Language Learning Aptitude

Let's take a specific example of how the theory of styles might be applied in one domain, that of learning foreign languages. Foreign-language learning in the United States is notoriously unsuccessful. The theory of styles suggests one reason why.

If different people learn best via different styles, then it may be that the kind of instruction that works for one person will not work for another. If we consider the discussion above and apply it to foreign-language learning, we can see in particular why many students will be frustrated: different people learn best in different ways. As a result, foreign-language learning aptitude isn't even a single construct. How well one will learn will depend in part upon aptitude, of course, but it will also depend on the match between the style of teaching and the style of learning.

A course that emphasizes memorization of vocabulary, memorization of grammar, and memorization of rote patterns is most likely to appeal to someone with an executive, local, conservative style. The person with this combination of styles prefers to be told what to do and how to do it—to be given material in multiple small doses, and to learn in traditional ways.

A course that emphasizes inductive audiolingual learning—expansion upon patterns presented orally, with little or no formal presentation of grammar—as is often found in courses bought in bookstores (and in the old FSI courses) is likely to appeal to an individual with who is somewhat legislative and somewhat executive, but not strongly either, because the course combines some inductive exploration and discovery with the presentation of large numbers of given patterns. Furthermore, though, the person will be local and conservative. Again, the language is built up in very small bits.

A course that emphasizes the direct method—learning from context—as is found in *French in Action*, *Destinos*, or *Español en Español*—is likely to be more appealing to the legislative, hierarchical, and liberal stylist. This course requires the student to construct the language for him or herself. Moreover, the individual is presented with large amounts of material, and needs to decide which are the important elements to be learned. The student is thrown into the new language and culture, and hence needs to be more receptive to thinking in a wholly new way—thus the benefit of the liberal style.

A course that strongly emphasizes comparison of the new language with the old, showing ways in which the new language is similar and ways in which it is different, is likely more to appeal to the judicial stylist. In this type of comparative course, which is more like the way classical languages are taught than it is like the way modern languages are taught, one essentially emphasizes translation: converting thinking in one language to thinking in the other. Thus, the new language is perceived in terms of the old, and every concept in the new language is compared, at least implicitly, to concepts in the old language.

Of course, the details of different courses will vary. The point to be made, however, is that an ideal foreign-language course will take into account not only an individual's abilities, but his or her styles as well.

Conclusions

Styles matter. Moreover, they are often confused with abilities, so that students or others are thought to be incompetent not because they are lacking in abilities, but because their styles of thinking do not match those of the people doing the assessments. Especially in teaching, we need to take into account students' styles of thinking if we hope to reach them.

We need carefully to consider how our practices in educational settings may deprive able people of opportunities, while giving opportunities to those who are less able. For example, extensive use of multiple-choice testing in the U.S. clearly benefits executive thinkers. Many tests of scholastic aptitude and other aptitudes confound measurements of styles with measurements of abilities. However, replacing all of these tests with projects and portfolios would simply result in a different group of students being benefited. Ideally, we need to teach and assess to a variety of styles.

The same principle applies in the world of work. Almost all jobs require an interview, but an interview, like any other form of assessment, tends to benefit people with certain styles at the expense of people with other styles. You will do better in an interview if you are external, and thus relate more readily and comfortably to your interviewer; are hierarchical, and can get into the interview the main points about yourself in a short amount of time; and if you are global enough to make sure that the interviewer gets the big picture about what you have to offer. This is not to say that there are no jobs for which these styles would not be beneficial. But these styles are not ideally suited to all jobs, so that the interview may be a better or worse selection device, depending on what it is being used for. Certainly, in the world of college admissions, it tends to favor a small subset of students over others who may be equally able.

Fortunately, some occupations allow flexibility in styles. For example, someone who wants to be a scholar might go into scientific research, which is more legislative, or into literary criticism, which is more judicial. Teachers may find themselves suited to the executive mode of many administrative jobs. Lawyers can become judges, giving themselves an opportunity to think in a more judicial way. So the world of work is sometimes tailored to allow people to express their stylistic preferences without changing career paths altogether. But such changes are not always possible, so that people need to think through what they do: an editor may be missing his or her chance to be a novelist, and vice versa.

So-called "gifted" adults are probably, in large part, those whose styles match their patterns of abilities. For example, someone with creative ability who has a legislative style will be at a distinctive advantage over someone lacking in creative ability who also has a legislative style. On the other hand, someone who is a strong analytic thinker may find a judicial style more suited to the ability than would be a legislative style. To succeed, you need to find compatibility between how you think and how you think well.

In sum, we need to take styles into account in the worlds of education and work, and the theory of mental self-government provides a way to do so. If we don't take styles into account, we risk sacrificing some of our best talent to our confused notions of what it means to be smart or a high achiever, when in fact some of the smartest people and highest achievers may only lack the style that we just happen to prefer.

Bibliography

Grigorenko, E., and Sternberg, R. J. (in press). Thinking styles. In D. Saklofske & M. Zeidner (Eds.), *International Handbook of Personality and Intelligence*. New York: Plenum, pp. 205-209.

Sternberg, R. J. (1988). Mental Self-Government: A Theory of Intellectual Styles and Their Development. *Human Development. 31*, pp. 197-224.

Sternberg, R. J. (1990). Thinking Styles: Keys to Understanding Student Performance. *Phi Delta Kappan. 71*, pp. 366-371.

Sternberg, R. J. (1993). Styles of the Mind. In *Restructuring Learning*. Washington, DC: Council of School Officers, pp. 21-31.

Sternberg, R. J. (1994). Thinking Styles: Theory and Assessment at the Interface between Intelligence and Personality. In R. J. Sternberg and P. Ruzgis (Eds.), *Intelligence and Personalities*. New York: Cambridge University Press, pp. 170-188.

Sternberg, R. J., and Grigorenko, E. L. (1993). Thinking styles and the gifted. *Roeper Review 16(2)*. pp. 122-130.

Sternberg, R. J., and Wagner, R. K. (1992). *Thinking Styles Inventory*. Unpublished test.

A Study of the Modern Language Aptitude Test For Predicting Learning Success and Advising Students

Madeline Ehrman
Foreign Service Institute, U.S. Department of State

The Modern Language Aptitude Test (MLAT) was part of a project examining biographical, motivational, attitudinal, personality, and cognitive aptitude variables among a total of 1,000 adult students preparing for overseas assignments at the Foreign Service Institute (with various smaller Ns for subsamples completing different instruments). Data were analyzed by correlation, ANOVA, chi-square, and multiple regression as appropriate to the data and the research questions. The MLAT proved the best of the available predictors of language learning success. As part of an effort to expand the concept of language learning aptitude beyond the strictly cognitive, this study relates the MLAT not only to end-of-training proficiency outcomes but also to personality dispositions, using both overall correlational data and information on extremely strong and weak learners. Qualitative findings from use of the MLAT part scores in student counseling activities are also described, suggesting utility for this well-established instrument beyond prediction of learning success.

This paper describes findings of research in progress at the Foreign Service Institute (FSI), a government language training institution. For years, incoming students have taken the Modern Language Aptitude Test (MLAT); indeed, a sample from FSI was among the groups on which the MLAT was originally normed (Carroll & Sapon, 1959). It is still in use as part of the agency's procedures for assignment to foreign language training. (Language aptitude testing is also done at other agencies.)

Over recent years the MLAT has become the subject of some controversy at FSI. Some program managers continue to see a good relationship between performance on the MLAT and in language training; others protest that the relation, such as it is, is not very strong and furthermore the MLAT may be not represent the true ability of those who lack formal education (Rockmaker, personal communication, 1993). Anti-MLAT opinion has also suggested that the MLAT was designed for the audio-lingual methodology that was in vogue in the late 1950's and 1960's and that the test is no longer valid for the much more "communicative" teaching that is now done at FSI (Bruhn, personal communication, 1992). Much of the distrust of the MLAT is doubtless connected with the increased suspicion of psychological testing during the last quarter century (Anastasi, 1988). The project on which this paper reports was initiated in order to take such questions about the MLAT out of the realm of allegation and find out just how useful it still is.

The present paper reports on two efforts to answer these questions. One is a quantitative investigation using a large sample of FSI students taken between 1992 and 1994. That study looks at the MLAT primarily as a predictor of language learning success in the FSI setting of intensive, full-time language learning for communicative use. The other portion of the paper describes a less rigorous attempt to make use of patterns of high and low MLAT part scores with individual students. The initial outcomes of this attempt, still highly exploratory, suggest that the

MLAT may have value for pinpointing areas of learning success and difficulty for a wide range of students, including some relatively able but context-dependent ones not well served by relatively grammar-oriented instruction.

Review of Literature

The MLAT was perhaps the culmination of a long tradition of psychometric test development and efforts to predict language learning achievement; and it achieved a fairly respectable level of success in the audio-lingual and grammar-translation classrooms of the 1950's and 1960's (Spolsky, 1995). Other important language aptitude tests developed out of the same tradition include the Pimsleur Language Aptitude Battery (Pimsleur, 1966) the Defense Language Aptitude Battery (Petersen & Al-Haik, 1976), and VORD (Parry & Child, 1990). The Pimsleur is different from the MLAT in particular because it includes a portion directly addressing the ability to infer language structure from an artificial language stimulus. The DLAB consists primarily of such induction-testing items, in a modified English. VORD was designed to test the ability to cope with the grammar of languages in the Altaic family and consists of items that test such grammatical prowess (Parry & Child, 1990). All four, including the MLAT, were found to have similar predictive validity (Parry & Child, 1990). This paper will not address these other instruments but will focus on the MLAT, which is the instrument that is still in use at the Department of State.[1]

The outcome of a major research project at Harvard University, the MLAT is based on a factor analysis of a large number of individual characteristics thought to contribute to language learning. Carroll (1962) describes the project in extensive detail; the MLAT Manual (Carroll & Sapon, 1959) provides information on the validation studies. The individual characteristics were grouped into four main categories: phonetic coding ability (distinguishing sounds and reflecting them graphically), grammatical sensitivity (recognizing and using syntactic relationships), memory (rote and contextualized), and inductive language learning. All but the last of these four are directly addressed in the five parts of the MLAT (see Figure 1).

Other components listed by scholars of language aptitude include motivation and knowledge of vocabulary in the native language (Pimsleur, 1968), the ability to hear under conditions of interference (Carroll, 1990), the ability to "handle decontextualizd language" (Skehan, 1991),and the ability to shift mental set and cope with the unfamiliar (Ehrman, 1994b, 1995, 1996; Ehrman & Oxford, 1995).

A desire for better prediction of language learning and the ability to exploit aptitude testing further has led to recent research efforts. At least two major projects in recent years have examined the role of individual differences in addition to strictly cognitive aptitude in language learning: the Defense Language Institute's Language Skill Change Project (Lett & O'Mara, 1990) and the Foreign Service Institute's Language Learning Profiles Project (Ehrman, 1993, 1994, 1995, 1996; Ehrman & Oxford, 1995; Oxford & Ehrman, 1995) investigated such variables as biographic factors, personality, motivation, anxiety, and learning strategies, as well general

[1] The remainder of the literature review owes much to a draft prepared by Frederick Jackson for an FSI roundtable at the Language Testing Research Colloquium in 1994 (Jackson, 1994).

intelligence (DLI only). A similar project has begun at the Central Intelligence Agency language school, though without personality variables, and DLI is engaged in a large-scale effort to improve the DLAB (Thain, 1992; Lett & Thain, 1994). This paper is part of one of these projects at FSI.[2]

Across a number of studies, predictive validity correlations for the MLAT have generally ranged between .42 and .62 for most languages, with outliers of .27 for certain non -Indo-European languages at the Defense Language Institute and as high as .73 for language instructor performance ratings at FSI (Carroll & Sapon, 1959). More recent tests of the MLAT are quite mixed. For instance, Brecht, Davidson and Ginsburg (1993) did not find the MLAT predictive of overall oral proficiency in intensive language training in Russia, though for the same programs they found Part III (Spelling Clues) to be "highly significant" in predicting listening comprehension and the Total Score to be significantly predictive of reading proficiency. They speculate that the lack of predictive value for oral proficiency is because this is a "communicative task." This suggestion is quite consistent with the questions raised at FSI (see above) and the point of view that standard aptitude measures do not "take into account" such developments as focus on communicative competence, pragmatics and discourse, new thinking by cognitive psychologists, etc. (Parry & Stansfield, 1990).

Another finding is that of Spolsky (1995), who reports that MLAT Part I correlated significantly with success on the part of Israeli learners of French as a foreign language, but the MLAT did not predict achievement in Hebrew at the same school, a variance he suggests may be related to differences in such factors as motivation, which is so powerful that it may override aptitude. (I suggest that it may also be the case that the students were learning Hebrew as a second language, not a foreign language, so not all their learning was classroom-based, which is the task for which existing language aptitude tests were designed.)

Most of the research cited addresses the use of the MLAT (and other aptitude measures) as predictors of learning success, and indeed this is an important consideration for assignment to intensive and long-term language training at taxpayer expense. However, a measure like the MLAT also has potential utility for *placement* in a program (Wesche, 1981) and *diagnosis* of learning difficulties, for counseling students, and for tailoring programs to their needs (e.g., Demuth & Smith, 1987; Sparks, Ganschow, & Patton, 1995). These applications have received far less attention in the literature. They are also among the areas of interest for the FSI investigation, and it is in these that the MLAT has been successfully used (Lefrancois & Sibiga, 1986; Wesche, 1981).

Methodology

Sample

In this study, there are 343 students altogether with at least a single Index score; of these, part scores for the subscales are available for 296. The mean age of the members of the sample is 37.

[2] The MLAT Project is separate but overlaps with the Language Learning Profiles Project, especially because for now it is using the same data set.

SD 9. Males constitute 59% and females 41% of the sample. The average age of students is 39, with a standard deviation of 9. The median education level is between bachelors and masters degrees. Of those that report previous language study, the average number of languages studied is 1.8. In the presentation of correlations with other instruments, *N*s are smaller, because not every person with an MLAT score in the data set completed all the other instruments.

FSI trains and tests students not only from its parent agency, but also from many other agencies. Student composition by agency and descriptions of student occupations in the sample at FSI would make identity of the institution obvious and is therefore omitted in this version.

Students in this study are beginners in long-term (i.e., 16 weeks or above) intensive language training. The languages they are studying are classified into four categories based on agency experience with the length of time needed by English speakers to reach "professional" proficiency (S-3 R-3—see 'Instrumentation' for a brief description of the ILR rating scale): 1. Western European; 2. Non-Western European but relatively quick for English speakers to learn (Swahili, Indonesian, and some North European languages); 3. Other non-Western European but excluding the category 4 languages (e.g., Russian, Thai); 4. "Super-hard" languages (Arabic, Chinese, Japanese, Korean)[3]. Usual training lengths vary by language category. Most FSI students are expected to reach "professional" proficiency (S-3 R-3) in 24 weeks in a category 1 language, in 32 weeks in a category 2 language, in 44 weeks in a category 3 language, and in 88 weeks (2 academic years) in a category 4 language.[4] These expectations are normally reflected in the lengths of student assignments to training and are also taken account of in the statistics reported in this paper.

Instrumentation

The MLAT. The Modern Language Aptitude Test (MLAT), (Carroll & Sapon, 1959) is the classic language aptitude test, with 146 items. The manual describes its five parts: I: number learning (memory, auditory alertness); II: phonetic script (association of sounds and symbols); III: spelling clues (English vocabulary, association of sounds and symbols); IV: words in sentences (grammatical structure in English); and V: paired associates (memorizing words), together with a total score. The MLAT was correlated .67 with the Primary Mental Abilities Test (Wesche, Edwards, & Wells, 1982), suggesting a strong general intelligence factor operating in the MLAT. Split-half reliabilities for the MLAT are .92–.97, depending on the grade or age. For college students, validity coefficients are .18–.69 for the long form of the MLAT and .21–.68 for the short form. For adult students in intensive language programs, validity coefficients are .27–.73 for the long form and .26–.69 for the short form (Carroll & Sapon, 1959). This study used the long form.

The subscales of the MLAT are described briefly in Figure 1. The Index Score used at FSI originated in the 1960's as a T-score based on the Total score, with three standard deviations of

[3] The Department of Defense uses a similar classification.

[4] Only three percent of students in this sample were studying category 2 languages—too small a number for most analyses. Category 2 and 3 languages are therefore combined for certain analyses.

10 on either side of a mean of 50.[5] It has since become frozen as a translation of the Total, much like Scholastic Aptitude Test ratings until recently, because of the agency personnel system's dependence on over 30 years of Index records. For users of the MLAT who are more familiar with the raw Total score, a table of equivalencies is provided in Appendix A.

Note that *Index 50 is the mean established when the MLAT was originally normed* and includes a variety of subjects from high schools and colleges. Whether it in fact is still representative of the population outside FSI is uncertain. What is certain, however, is that a mean Index of 50 is no longer valid for FSI students. There has been a gradual upward tendency in the MLAT Index mean at FSI over the intervening 30 years: Wilds (1965) reported a mean Index of 54 (N=957, no SD); an agency-internal document reports a 1984 mean Index of 59 , SD 10, N-312 (Adams, 1984); and the mean Index for all the students in the current sample who had MLAT scores is 63 SD 10, N = 343.[6]

Figure 1. MLAT Subscales

Part I. Number Learning: This subtest requires the examinee to learn four morphemes and interpret them in combinations that form numbers; it is entirely orally delivered. The subtest is described in the Manual (Carroll & Sapon, 1959) as measuring part of memory and "auditory alertness" which play a part in auditory comprehension (showing how well one understands what one hears) of a foreign language.

Part II. Phonetic Script: This subtest requires the examinee to select a written equivalent (in Trager-Smith phonemic transcription) for an orally delivered stimulus. The MLAT Manual describes the subtest as dealing with the ability to associate a sound with a particular symbol, as well as how well one can remember speech sounds. In addition, the subtest is described as tending to correlate with the ability to mimic speech sounds and sound combinations in a foreign language.

Part III. Spelling Clues: In this entirely written subtest, an English word is presented in a very non-standard spelling. The examinee must select the correct synonym. Vocabulary items are progressively more difficult, though the most difficult is probably within the repertoire of a college graduate. According to the Manual, scores on this part depend largely on how extensive a student's English vocabulary is. As in Part II, it measures the ability to make sound-symbol associations but to a lesser degree.

Part IV. Words in Sentences: The stimulus is a sentence with an error. The examinee must indicate which part of another sentence matches the designated part. The subtest is entirely in writing. It is described as dealing with the examinee's sensitivity to grammatical structure and thus expected to provide information about the ability to handle grammar in a foreign language. No grammatical terminology is used, so scores do not depend on specific memory for grammatical terms.

[5] Although Appendix A lists possible Index Scores below 20, current scoring devices do not yield Index Scores below 20.

[6] The MLAT was standardized in part on an FSI sample. Although that sample, as a result of the times (late 1950s) was all male, no gender differences have appeared on the MLAT among present students on any subtest of the MLAT or on its Total or standardized score.

Part V. Paired Associates: The examinee is presented with 24 foreign words with their English equivalents and given some time to learn them. The words are then tested. This subtest is said to measure the examinee's ability to memorize by rote--a useful skill in learning new vocabulary in a foreign language.

Raw Score Total. Total of all five subscales.

Index Score. Originally a scaled (T) score used at FSI that is based on the Total. The original mean was 50, with a standard deviation of 10. These norms are now out of date; the Index is now simply a conversion of the raw Total into a scale ranging between 20 and 80. Local norms using the Index have not been formally established because the Index score using the original norms is deeply embedded in the agency's personnel system.

End-of-training proficiency tests. These tests provide the main criterion measure in this study. At the end of training, FSI students are given proficiency assessments resulting in ratings ranging from 0 to 5 for speaking (the S-score, which includes interactive listening comprehension) and for reading (the R-score). The full oral interview, including speaking, interactive listening, and an interactive reading test using authentic materials, takes two hours. R-3, for example, indicates reading proficiency level 3 ("professional" proficiency); S-2 represents speaking proficiency level 2 (working proficiency). Other levels are 0 (no proficiency), 1 (survival level), 4 (full professional proficiency, with few if any limitations on the person's ability to function in the language and culture), and 5 (equivalent to an educated native speaker).

The ratings are equivalent to the guidelines of the Interagency Language Roundtable/American Council on the Teaching of Foreign Languages (ILR/ACTFL) that originated at FSI and have been developed over the years by government agencies. (These guidelines are detailed by Omaggio, 1986). Most students enter FSI with goals of end-training proficiency ratings at S-3 R-3 for full-time training, comparable to ILR/ACTFL Advanced Proficiency.

Reliability studies have shown that government agencies have high interrupter reliability for proficiency ratings within a given agency, but that the standards are not always the same at every agency; thus raters at different government agencies do not have as high an interrater reliability as raters at the same agency. Proficiency ratings are thus considered reliable indicators of the level of language performance of an individual student within an agency (Clark, 1986. "Plus" scores (e.g., indicating proficiency between S-2 and S-3) were coded as 0.5; thus, for example, a score of S-2+ was coded 2.5.

Learning style, strategy, and personality instruments. The Learning Style Profile is a pure learning style instrument. The Myers-Briggs Type Indicator and its Type Differentiation Indicator scoring system is both a personality instrument and a way to assess learning style, as is the Hartmann Boundary Questionnaire. The student learning activities questionnaires tap learning strategies.

The Hartmann Boundary Questionnaire (HBQ) (Hartmann, 1991). The HBQ was developed for research with sleep disorders and nightmares, using a psychoanalytic theoretical base. It is intended to examine the degree to which individuals separate aspects of their mental, interpersonal, and external experience through "thick" or "thin" psychological boundaries. Its 146 items address the following dimensions: sleep/dreams/ wakefulness, unusual experiences,

boundaries among thoughts/feelings/moods, impressions of childhood/adolescence/adulthood, interpersonal distance/openness/ closeness, physical and emotional sensitivity, preference for neatness, preference for clear lines, opinions about children/adolescents/adults, opinions about lines of authority, opinions about boundaries among groups/peoples/nations, opinions about abstract concepts, plus a total score for all twelve of the above scales. Hartmann found women and younger people to score consistently "thinner" than men and older people. Cronbach alpha reliability for the HBQ is .93, and theta reliabilities for subscales are .57–.92 (Hartmann, 1991).

The National Association of Secondary Schools Principals' Learning Style Profile (LSP), (Keefe & Monk, with Letteri, Languis, & Dunn, 1989). This is a 125-item composite measure composed of many different approaches to measuring learning style. The main subscales are cognitive skills (analytic, spatial, categorization, sequential processing, detail memory, discrimination), perceptual response (i.e., sensory preferences: visual, auditory, emotive/kinesthetic), orientations (persistence, verbal risk-taking, manipulative), study time preferences (early morning, late morning, afternoon, evening), and environmental context for learning (verbal vs. spatial, posture, light, temperature, mobility, and grouping). Cronbach's alpha for the subscales ranged from .47 to .76, with an average of .61. Test-retest reliabilities were .36 to .82 after 10 days and somewhat lower after 30 days. Concurrent validity of the LSP's analytic subscale with the Group Embedded Figures Test was .39. Concurrent validity of the perceptual response subscales of the LSP with the Edmonds Learning Style Identification Exercise was .51 - .64. Many of the environmental context subscales of the LSP correlated with Dunn and Dunn's Learning Style Inventory, .23 - .71. All concurrent validity scores are reported in the manual with a significance value < .002.

The Myers-Briggs Type Indicator (MBTI), (Myers & McCaulley, 1985) Form G. This instrument is a 126-item, forced-choice, normative, self-report questionnaire designed to reveal basic personality preferences on four scales: extraversion-introversion (whether the person obtains energy externally or internally), sensing-intuition (whether the person is concrete/sequential or abstract/random); thinking-feeling (whether the person makes decisions based on objective logic or subjective values); and judging-perceiving (whether the person needs rapid closure or prefers a flexible life). Internal consistency split-half reliabilities average .87, and test-retest reliabilities are .70 - .85 (Myers & McCaulley, 1985). Concurrent validity is documented with personality, vocational preference, educational style, and management style (.40 - .77). Construct validity is supported by many studies of occupational preferences and creativity.

The Type Differentiation Indicator (TDI) (Saunders, 1989). The TDI is a scoring system for a longer and more intricate 290-item form (MBTI Form J) that provides data on the following subscales for each of the four MBTI dimensions: extraversion-introversion (gregarious-intimate, enthusiastic-quiet, initiator-receptor, expressive-contained, auditory-visual); sensing-intuition (concrete-abstract, realistic-imaginative, pragmatic-intellectual, experiential-theoretical, traditional-original); thinking-feeling (critical-accepting, tough-tender, questioning-accommodating, reasonable-compassionate, logical-affective); and judging-perceiving (stress avoider-polyactive, systematic-casual, scheduled-spontaneous, planful-open-ended, methodical-emergent). The TDI includes seven additional scales indicating a sense of overall comfort and confidence versus discomfort and anxiety (guarded-optimistic, defiant-compliant, carefree-worried, decisive-ambivalent, intrepid-inhibited, leader-follower, proactive-distractible), plus a

composite of these called "strain."7 Each of these comfort-discomfort subscales also loads on one of the four type dimensions, e.g., proactive-distractible is also a judging-perceiving subscale. There are also scales for type-scale consistency and comfort-scale consistency. Reliability of 23 of the 27 TDI subscales is greater than .50, an acceptable result given the brevity of the subscales (Saunders, 1989).

Student Learning Activities Questionnaires. At the end of training, each student in the study was asked to complete two questionnaires: "CLASSACT" (Ehrman & Jackson, 1992) on relative usefulness of a fairly detailed list of classroom activities (Likert scaled 1-3) and "SELFACT" (Hart-Gonzalez and Ehrman, 1992) on relative usefulness (1-3) of their own study activities and estimated time per week devoted to each. These questionnaires are used here for the first time. Because completion at end of training was voluntary and students were very busy with preparations for departure, the return rate was low, and *N*'s for a number of the items are not adequate for analysis. This and other studies using these two questionnaires are part of their validation. When there are sufficient cases, they will be subjected to reliability analysis and factor analysis.

Data Collection and Analysis

Data collection took place over a two-year period, between 1992 and 1994. Students at each of the two annual major intakes were asked to participate but could decline the invitation; under 5% of the students who were approached chose not to participate. During the 1992-1993 academic year, all French and Spanish students (who start 10 times a year) were also invited to join the study, with the same drop-out rate.

All questionnaires except the MLAT were administered within the first week of training. If a student already had an MLAT record, he or she could arrange for those scores to be included in the research data set; otherwise, MLAT administration took place within the first month of the beginning of training. In this sample, almost all (95%) of the MLAT scores were current, i.e., within the previous 3 years. Proficiency tests were administered at the end of training, after (in most cases) 24 or 44 weeks.

Data analysis in this study on SPSS for Windows 5.0.1 (Norusis, 1992) used correlations, one-way analysis of variance (ANOVA), t-tests, and multiple regression. Correlations of the MLAT were done with end-of-training ratings for speaking and reading proficiency (the FSI proficiency test is described above, under "Instrumentation") and with individual difference variables (see above for listing and descriptions of the instruments). The data used for the correlations between end-of-training proficiency and the MLAT Index for all language categories combined were filtered to equalize expected length of training and proficiency outcomes (to make results of a language like French comparable to those of a language like Chinese).

RESULTS

Distributions

Table 1 shows that the Index Score is somewhat higher for category 2, 3, and 4 languages than for category 1 languages in central tendency and range (see "Sample" for definitions of these categories). The part scores follow the same pattern.

Table 1: MLAT Descriptive Statistics for Index Score

Category	*N*	Mean	SD	Range	Mode	Skewedness	Kurtosis
All Students	343	63	10	21 - 80	70	-.973	1.392
Category 1	169	59	12	21 - 80	61, 70	-.808	.625
Categories 2-3	120	66	8	45 - 80	70	-.462	-.171
Category 4	54	63	10	26 - 78	64	-.900	.770

Minimum possible Index: 20; maximum possible Index: 80.
Category 1: Western European languages; Category 2: Swahili, Indonesian, Malay; Category 3: Eastern European and non-Western languages (except Category 4 languages); Category 4: Arabic, Chinese, Japanese, Korean.

The distributions, with their high central tendencies and reduced space below the ceiling for FSI students, reflect several forms of preselection. The first is that many students have self-selected for foreign affairs careers. Most of these went through their agency's selection process. This process has already probably eliminated some of the students least likely to score well on the MLAT. Second, the MLAT Index Score is used for selection of students in the FSI's parent agency's personnel system, along with other evidence of likely learning, especially evidence of previous language learning success. (Such selection is authorized in the personnel regulations for U.S. Department of State, though it is clearly stated that evidence of learning success overrides the MLAT.)

Selection is done in the State Department's personnel system especially for non-Western-European languages, for which training to the "professional" proficiency level (S-3 R-3) takes 44-88 weeks. Relatively low MLAT students (Index below 55 for category 3 or 60 for category 4 languages) with no other evidence of success are normally sent to Western European languages by preference, hence this is where we find a relatively large range of tested aptitude.

The effect of preselection using the MLAT for category 3 and 4 languages is to make it very difficult to analyze the MLAT's predictive value for these languages in this sample. On the other hand, in view of the expense entailed by 44-week and 88-week intensive language training, assignments personnel understandably seek every indication of likely success or lack of it, without reference to the needs of the researcher.

Other results are described under two rubrics: findings related to prediction of language learning success and findings related to diagnosis and student counseling. The former are quantitative; the latter are qualitative.

Results related to prediction of language learning success

Correlations: Correlation coefficients for MLAT Index, Total, and part scores with S- and R-ratings range in the 40's and 50's for the MLAT when a broad range of scores is available, comparable with coefficients found originally by Carroll (1990). The Index Score tends to show higher correlations with end-of-training proficiency ratings than the part scores or the Total. Correlations for the Index Score are shown in Table 2.

Table 2: Correlations of MLAT Index Score with End-of-Training Proficiency Ratings

	r	S-rating	r	R-rating
All languages:	.44	(N = 343)	.40	(N = 341)
Category 1 languages:	.52	(N = 169)	.55	(N = 168).
Category 2-3 languages	.34	(N = 120)	.35	(N = 120)
Category 4 languages	.47	(N = 54)	.34	(N = 53)

Category 1: Western European languages; Category 2: Swahili, Indonesian, Malay; Category 3: Eastern European and non-Western languages (except Category 4 languages); Category 4: Arabic, Chinese, Japanese, Korean. S-rating: speaking and interactive listening; R-rating: reading.

Correlations are weakest for category 2-3 languages and strongest for category 1 languages, where there is the greatest range and the distribution of MLAT scores closely resembles a normal distribution. For categories 1-3, correlations with reading and speaking are roughly the same. In category 4 languages, they are stronger for speaking than for reading. This difference may be because there is less range in reading scores (they are much lower for beginners than in other languages), or possibly because the MLAT does not address abilities needed for reading languages that use Chinese characters--three out of the four category 4 languages.

T-tests: Cut points were established such that the cut was made between a score and all those below it. For example, a cut point of S-2 divides between cases less than S-2 and those equal to or greater than S-2. T-tests were done at each cut point from 1+ to 3+ (there were not enough 4-level scores in the sample for meaningful statistics).[7] P-values range from .0001 to .044; with a few exceptions as indicated, only those at the .0001 level are reported in Table 3.

Table 3. Cut points at which the part discriminates at .0001 significance

MLAT Part	ILR Speaking Level	ILR Reading Level
Part I	S-2, S-3	
Part II	S-2	R-2
Part III	S-2, S-2+, S-3	R-2, R-2+, R-3
Part IV	S-2, S-3	R-2
Part V	S-2	
Total Score	S-1+, S-2, S-3, S-3+[8]	R-2, R-2+, R-3, R=3+
Index Score	S-1+, S-2, S-3, S-3+[8]	R-2, R-2+, R-3, R=3+

[7] A table of the T-test results is available on request.

[8] The Total Score discriminates at the S-3+ cut point at a significance of .012, and the Index Score at the .013 level.

The best discriminators at all levels of proficiency appear to be Parts III (English vocabulary in altered spellings) and the Total and Index Scores. Another summary of the same results appears in Table 4, this time organized by cut point:

Table 4. Part(s) discriminating at .0001 significance at each ILR score cutpoint

Speaking:	S-1+	Total Score
	S-2	Parts I, II, III, IV, V Total Score, and Index Score.
	S-2+	Part III, Total Score, and Index Score
	S-3	Parts I, III and IV, Total Score, and the Index Score
	3+	The Total and the Index Scores[8]
Reading:	R-1+	None
	R-2	Parts II ,III, and the Total Score
	R-2+	Part III and the Total Score.
	R-3	Part III, the Total Score, and Index Score
	R-3+	The Total Score and the Index Score.

Analysis of Variance: This investigation was done only for the entire sample, because the numbers of subjects were not sufficient for category 2–3 or 4 languages separately. In a study of the extremely strong and weak students in the sample, the bottom 3–4 percent were contrasted against all others and top 5–6 percent against all others. Extreme students were selected on a formula that combined length of training, relative difficulty of language by category, and end-of-training scores. There were fewer students at the low end because the very weakest may be withdrawn well before scheduled end of training and because both teachers and students make every effort to reach the student's training goal, which in most cases is S-3 R-3. More detail on the extremes study, including the selection formula, is available in Ehrman (1994b).

Data for the individual difference variables were analyzed using the one-way analysis of variance procedure in SPSS for Windows 6.1. The findings for the MLAT are displayed in Tables 5 and 6.

Speaking: Of all the variables analyzed, the Parts III, IV, V, the Total, and the Index scores best differentiated the *weakest* students, that is, these variables had the largest F-scores. The MLAT variables also differentiated these weak students better than any other of the many variables in the study.

For the *strongest* students' speaking scores, the Index ($F=7.83$, $p < .0055$) was the strongest differentiator from among the MLAT and learning style variables, but it was not as good as these biographical background variables: education level, number of previous languages, and previous highest score in speaking and especially reading. The MLAT appears to differentiate the strongest speakers less clearly than the weakest speakers and readers and the strongest readers.

Table 5: Speaking Performance Extremes: ANOVAs

Weakest, Speaking		N selected (weakest): 4 (Parts & Total), 6 (Index) N not selected (all others)= 292 (Parts & Total), 337 (Index).				
Part	**Weakest Mean**	**All Others Mean**	**Weakest SD**	**All Others SD**	**F**	**sig.**
I	24.5	36.5	6.5	9.1	6.8524	.0093
II	18.5	24.7	3.5	4.5	7.3634	.0070
III	11.0	28.3	8.6	9.9	12.1415	.0006
IV	15.3	28.0	5.3	7.5	11.4289	.0008
V	11.5	19.3	4.7	5.3	11.4289	.0008
Total	80.8	136.7	24.6	27.5	16.3881	.0001
Index	43.2	62.7	10.8	10.5	20.5548	.0000

Strongest, Speaking		N selected (strongest): 14 (Parts & Total), 19 (Index) N not selected (all others) = 281 (Parts & Total), 324 (Index).				
Part	**Weakest Mean**	**All Others Mean**	**Weakest SD**	**All Others SD**	**F**	**sig.**
I	40.5	35.0	4.9	9.7	4.4395	.0362
II	27.1	24.3	2.8	4.7	5.2765	.0225
III	32.8	27.0	7.0	14.2	4.5701	.0336
IV	30.0	27.2	5.0	7.9	1.7067	.1927 ns
V	20.8	18.8	4.2	5.5	1.6950	.1942 ns
Total	151.2	132.5	13.8	29.6	5.7291	.0175
Index	68.2	60.9	5.9	11.2	7.8286	.0055

Data analysis done by SPSS for Windows v. 6.1, One Way Analysis of Variance Test. Degrees of freedom are available on request.

Reading: For reading, Parts III and IV and the Total and Index Scores best differentiate the *weakest* students. The *strongest* are differentiated clearly by all MLAT parts except Part IV; with the Index Score providing the clearest distinction.

Table 6: Reading Performance Extremes: ANOVAs

Weakest, Reading		N selected (weakest): 3 (Parts & Total), 4 (Index) N not selected (all others) = 292 (Parts & Total), 337 (Index).				
Part	**Weakest Mean**	**All Others Mean**	**Weakest SD**	**All Others SD**	**F**	**sig.**
I	23.0	36.4	7.0	9.1	6.4559	.0115
II	17.7	24.7	3.8	4.5	7.1481	.0079
III	7.3	28.2	5.5	9.9	13.4109	.0003
IV	13.0	28.0	3.5	7.5	11.8901	.0006
V	11.0	19.3	5.6	5.3	7.3757	.0070
Total	72.0	136.6	21.2	27.6	16.3758	.0001
Index	40.5	62.7	12.6	10.5	17.6391	.0000

Strongest, Reading		N selected (strongest): 78 (Parts & Total), 93 (Index) N not selected (all others)= 217 (Parts & Total), 248 (Index).				
Part	Weakest Mean	All Others Mean	Weakest SD	All Others SD	F	sig.
I	38.9	33.8	6.3	10.5	15.0647	.0001
II	26.1	23.7	3.5	4.8	15.4653	.0001
III	31.0	26.9	8.6	10.2	14.7692	.0002
IV	29.2	26.7	6.5	7.9	6.1293	.0140
V	21.3	17.9	4.1	5.6	22.5703	.0000
Total	146.5	128.0	20.9	30.1	23.7211	.0000
Index	66.3	59.6	8.0	11.3	26.1914	.0000

Data analysis done by SPSS for Windows v. 6.1. One Way Analysis of Variance Test. Degrees of freedom are available on request.

Multiple Regression: Multiple regression analysis for end-of-training speaking and reading examined the effects of age, education level, number of previous languages studied, highest previous speaking and reading ratings, a general motivation rating, two self-efficacy ratings (self-rated aptitude and expectation of success in this course), two anxiety ratings (for the course in general and about speaking in class), and the MLAT Index Score.

For speaking, the analysis yielded a multiple R of .40, R Square of .16, with two predictors in the equation: the MLAT Index Score (Beta .32, T = 3.293 p = .0014) and Highest Previous Reading Score (Beta .21, T = 2.208, p = .0297).

For reading, the analysis yielded a multiple R of .37, R Square of .14, with the same two predictors in the equation: the MLAT Index Score (Beta .27, T = 2.798, p = .0063) and Highest Previous Reading Score (Beta .22, T = 2.266, p = .0258).

Results related to diagnosis and student counseling

In this section, both quantitative and qualitative findings are described, as part of an ongoing effort to build learner profiles that can be used by teachers, teacher trainers, program managers, and even students themselves to enhance student learning. The quantitative results contribute to a fuller picture of the kinds of students who are advantaged and disadvantaged in full-time intensive and largely communicative language training, by adding personality factors to more cognitive abilities. The qualitative material is very exploratory, but it has been promising enough to merit description here so that others can use and test the emerging patterns. It is also included here because it provides more information on what the MLAT may actually be measuring, and because it sheds more light on the complexity of the apparently simple factor-analysis-based MLAT parts.

Relationships with Other Individual Difference Variables: There are other variables than the MLAT that are useful in the building of an individual learner profile that can be used for diagnosis and counseling (the utility of these for prediction is more directly addressed in Ehrman, 1993, 1994a, b, 1995, 1996, Ehrman & Oxford, 1995; Oxford & Ehrman, 1995). These variables bear interesting relationships to the MLAT. Correlations of at least .30 between the MLAT Index

Score and/or Total Score and other instruments used in the larger study are presented in Table 7. The correlations suggest the relationships described below.

Table 7: MLAT Index or Total Score Correlations with Other Variables

Variable	Lang. Category Grp	rho	Correlate N	
Number of Previous Langs	All	.40**	Index	245
HBQ Prefer Blurred Edges	Cat. 1	.51*	Total	25
HBQ Prefer Low Neatness	Cat. 1	.47	Total	25
HBQ Thin External Boundaries	All	.32**	Total	102
HBQ Total Score (thin)	All	.30**	Index	110
MBTI/TDI Intellectual (N)	Cat. 1	.45**	Index	96
MBTI/TDI Intellectual (N)	Cat. 2-3	.35**	Index	103
MBTI Intuition	Cat. 1	.34**	Total	93
MBTI Imaginative (N)	Cat. 1	.34**	Index	96
MBTI Introversion	Cat. 1	.30*	Total	93
LSP Simultaneous Processing	Cat. 1	.45	Index	24
LSP Sequential Processing	Cat. 1	.43	Index	24

*All the above correlations are significant at least at the .05 level; * indicates the .01 level; ** indicates the .001 level. HBQ: Hartmann Boundary Questionnaire. MBTI: Myers-Briggs Type Indicator. LSP: Learning Style Profile. "Imaginative" and "Intellectual" represent the intuition (N) poles of the MBTI TDI Realistic-Imaginative and Pragmatic-Intellectual subscales for the sensing-intuition main scale.*

Those who have scored high on the MLAT tend to have studied languages previously, often prefer an "intuitive" approach to taking in information on the MBTI. MBTI intuition indicates preferences for the abstract over the concrete, search for meaning, a preference for the "big picture" rather than details, and the speculative over the strictly experiential (Myers & McCaulley, 1985). They describe themselves as having relatively thin ego boundaries, especially with respect to such matters as dislike for too much neatness, order, and clear-cut separations among visual images. Thin ego boundaries, correlated with MBTI intuition, indicate receptivity to a wide range of experience, both internal and external, and a willingness to blur categories. This concept is used to operationalize a model of tolerance of ambiguity (Ehrman, 1993). High-MLAT students also are often more skilled at simultaneous and sequential visual processing on the LSP.

The analyses of variance in the extremes study support these findings for extremely strong and weak students and add as an advantage a preference for a flexible approach shown in the perceiving pole of one of the MBTI/TDI JP subscales, methodical vs. emergent. (This subscale of the TDI scoring of the long MBTI opposes a desire to know in advance what will happen in contrast with a preference to let events "emerge" and cope with them as they come up; the strongest students indicated a preference for an emergent approach.)

The MLAT and Learning Activities. A recent correlation study showed interesting relationships between the MLAT and a set of activities that students rated for perceived utility both before starting training and at the end of training (Ehrman, 1995). The correlations were

similar for both pre- and post-testing. Though the correlations were generally low (mostly 20's and some in the 30's), there seemed to be suggestive patterns in them when subjected to a content analysis. Findings described below were based on the content analysis of those items with which the MLAT was correlated and on correlations of MLAT scales with variables from the other instruments (Table 7).

In summary, high MLAT Index and *all* part scores correlate with items that are interpreted as reflecting self-confidence as a language learner and tolerance of ambiguity (low-structure activities and input).

The Index and parts II, III, IV, and V are correlated with items suggesting acceptance of/preference for use of authentic material for reading and listening and authentic conversation.

Parts III and IV are correlated with items suggesting endorsement of learning activities that reflect an analytic, structured approach. This effect was slightly stronger for part III; students who rejected a "touchy feely" approach on one item (the only such item) also tended to be high scorers on part III.

In contrast, a more experiential, kinesthetic approach may be suggested by the Index and a peak on part II, at least as indicated by the correlations with preferred learning activities.

Students who endorsed activities that were interpreted as indicating a preference for directing their own study tended to do well on the Index and parts II and IV.

Interpreting part-score profiles. The above patterns suggested possible uses for the MLAT profile in student counseling, where they currently being tested. Some profiles that these data suggest are outlined in below.

1. All parts high (a very high Index will usually represent this kind of profile):
   has done well on all the parts
   self-confident as learners
   respond well to activities that require tolerance of ambiguity
   like relatively unstructured learning
   enjoy and even prefer authentic input.

A related analysis found a relationship between endorsement of relatively unstructured, ambiguous, authentic activities and higher end-of-training scores (Ehrman, 1995).

2. A more uneven profile in which parts III (especially) and IV are high:
   analytic learner, perhaps field independent
   likes a program with a clear plan (<u>not</u> the same as a restrictively sequential program).
   usually has good knowledge of English vocabulary and grammar.
3. An uneven profile in which Part II is highest, together with a strong Index, (most other parts above average) may indicate a student who likes experiential, hands-on, participatory learning.

4. An uneven profile in which Parts II and IV relatively high, together with a strong Index, may suggest a student who likes to take control of his or her own learning sequence and can use both analytic and global learning strategies comfortably.
5. When either part I or part V is the highest of the part scores, there so far seems to be little that is distinctive, though interviews are suggesting that low scores on part V appear to indicate either poor mnemonic skills or weak metacognitive strategies, or both.
6. All parts low (a very low Index will usually represent this kind of profile):
   has done poorly on all the parts
   often lacks self-confidence as a learner and subject to anxiety because of slow progress
   likely to be overwhelmed by unstructured and uncontrolled input
   will need a great deal of scaffolding for longer than most other students
   likely to progress slowly

Overall Total score on the MLAT or the Index gives a crude measure useful when it is either very low or very high: a very low Total or Index score indicates weakness in all the factors; a very high score suggests strength in all the factors. When the Index falls in the middle range--roughly within a standard deviation of the mean-- it becomes much more important to examine the "scatter" of the part scores.

Using part scores with students. The student counseling activity uses the variations in part scores to initiate interpretations that are raised with the student to examine how he or she learns. Interpretation usually requires an interview of the student. Responses by students to the question "What happened when you were doing this part?" provides useful information about the skills tested by each part. Each of the MLAT factors probably represents a set of abilities. For example, Part III has proved particularly fruitful in the diagnostic process with students. Among the possible task requirements of this item are: gestalt processing of the whole word; sound-symbol processing; rapid hypothesis testing of sound-symbol possibilities; shift in mental set; and semantic evaluation.

These task requirement possibilities are represented as student performance in the following six cases of poor outcome on Part III, each of which is followed by implications for the classroom. The cases represent composites of responses actually received to the query about what happened while students were completing this subtest. (Many examples of real cases with specific score profiles, are to be found in Ehrman, 1996.)

1) One student might have done poorly on Part III because of difficulty with the kinds of analytic activities often described as "field independent." This student is likely to have difficulty with induction of rules and patterns and with grammar-oriented activities that have little context. Students of this sort usually find more contextual learning helpful.

2) Another might do poorly on the same part because of a weak English vocabulary (among the possible causal factors: poor education, low intelligence). This student, if a native speaker of

English,[9] may have difficulty with vocabulary learning (among other things) because of lacking concepts and background. The classroom may have to include activities to help this student build content background as well as language.

3) A third experiences difficulties reorganizing schemata or with gestalt processing or shifting mental set. Part III makes considerable demands on a person's ability to shift mental set. Such a student may be more comfortable with relatively predictable activities and less so with open-ended ones and may need assistance in building skills for coping with the unfamiliar or unexpected.

4) Yet another might have a *phonetic coding difficulty* of the sort described by Sparks, Ganschow et al. (1991), i.e., working with sound-symbol relationships. He or she is likely to have corresponding low scores in Parts I and II, which also require decoding of sounds. Such a student is likely to be handicapped in both speaking and reading and will need more time to absorb material. Kinesthetic input is likely to help this student.

5) Links among extraversion, desire for language use outside the classroom, and MLAT Part III suggest a *distractibility* factor. That is, a strongly extraverted student who is drawn to interpersonal interactions might not be as adept at the kind of focus that the puzzle solving aspect of Part III entails as one who tunes out the world more readily. Study strategies, including frequent breaks and setting up conditions to maximize concentration, might help a student who has difficulty concentrating.

6) Finally, a person who is reminded by Part III items of crossword puzzles and dislikes them has had an *affective* reaction which interferes with ability to use cognitive resources. Alternatives to "puzzle-solving" activities would probably help the sixth student, or perhaps cooperative learning when puzzle-like activities are part of the curriculum. The teacher would need to be alert to the affective impact of these activities.

Interpretation of a student's profile is made more complex by factors that can affect any or all of the parts of the test. In some cases, a low score on Part III (or any other part) may be the result of a mechanical error, such as marking in the wrong row of the answer sheet. sometimes a student will say that he or she did not understand the instructions for a given part. (This response raises questions about attention, motivation, or test-taking strategies.) Some students ascribe low scores to fatigue, which is plausible especially for the later parts. Interpretation is further complicated by the fact that a student might suffer from several of these difficulties at once.

DISCUSSION

Summary: Despite the effects of restricted range, skewed distribution, and relatively limited ceiling (because of negative skew for this high-end sample), the MLAT remains the best predictor of the variables examined. In general, the Index Score is the most useful of the MLAT variables

[9] The MLAT is designed for use with native speakers of English. At FSI it is considered invalid for non-native speakers, though if one takes it and does well (Index greater than 50), such performance is considered a promising sign. Low scores, on the other hand, are ignored.

as a predictor (strong in all cases, and with highest correlation coefficients). Of the part scores, Part III is the strongest predictor. Part III, with its dependence on knowledge of English vocabulary as well as ability to solve puzzles, may also be an indirect indicator of general intelligence. This would apply to both fluid ability, because of the cognitive restructuring required by the task, and crystallized ability (vocabulary), and "g" or general intelligence, since general vocabulary is also considered to be the single best stand-in for overall intelligence (Anastasi, 1988, Wesche, Edwards, and Wells, 1982).

Is the MLAT more suitable for Western European languages than for non-Western languages? The question remains open. Correlations and T-tests show stronger results for category 1 languages than for 2, 3, and 4 languages. On the other hand, the substantial preselection of students suggested by the very skewed distribution and the restriction of range in the sample may account for this finding as much as appropriateness of the MLAT for non-European languages. Furthermore, the fact that the correlations for category 4 language outcomes are actually better than those for category 3 languages, despite substantial truncation of range, might suggest that the MLAT is actually a fairly strong predictor for these languages. (The higher correlations might also be related to the much smaller *N* for category 4 languages.) We cannot test either hypothesis on the FSI language-student population as long as they are preselected and preselected using the MLAT.

Of the extended set of variables in the research project (including learning strategies, cognitive styles, motivation, anxiety, and personality variables), the MLAT Index Score also continues to be the strongest, both in the correlation coefficients and ANOVAs of extremely weak and strong students. It is especially powerful as a selector of extremes.

In addition to the relatively crude information provided by the Index score that may help in selection for training, the part-score profile shows promise as a way to better target classroom interventions and advice to students about appropriate learning strategies to develop. High performance on the MLAT appears to be related to personality variables that indicate high tolerance for ambiguity and the ability to reconceptualize input.

Is the MLAT passé in an age of communicative teaching? The MLAT has been criticized by many as rating aptitude only for audio-lingual training, which was in vogue when the MLAT was developed. However, the MLAT correlations remain about the same, even though the teaching methodology has changed considerably (most FSI courses now have a substantial communicative component, and some are almost wholly communicative). Why is this so? The following are some possibilities.

1. Perhaps the MLAT is really multidimensional, and a different set of dimensions applies to different methodology.
2. Perhaps the operative factor is really some form of coping with ambiguity or coping with the unfamiliar.
3. Possibly it is the "g" (general intelligence)-factor that is operative for FSI students. (Sasaki (1993) found a general found a general cognition factor, which she describes as similar to

"g," to account for 42% of the variance among Japanese college students studying English as a foreign language)

4. The very nature of classroom training may make a difference. Although FSI classroom training requires the ability to cope with communicative activities and access global and inferential learning, it also makes heavy demands on analytic skills. These may become increasingly important at higher proficiency levels; this fact may be why part III, which is most strongly associated with analytic learning, differentiates most at the higher levels in the T-tests and why parts III and IV together are the most predictive of extremes in achievement, together with the Index, which is more associated with predilection for the more open-ended learning that is also necessary for achieving high proficiency levels in FSI classrooms. The study of ego boundaries using the Hartmann Boundary Questionnaire (Ehrman, 1993) found a similar construct, labeled "tolerance of ambiguity" to be essential to effective classroom learning at FSI. In this study, thin ego boundaries that let a student take in new data were not enough alone--students had to impose some sort of mental structure on their intake and at the same time stay open to the fact that their structures were hypothetical. Investigation now under way is examining the applicability of the field independence construct to these findings, further information on which is to be found in Ehrman, 1996.

The aptitude concept: Expanding the aptitude concept is one of the subjects of an ongoing investigation of individual differences in language learning. The subject is discussed in greater detail in Ehrman, 1994b, 1995, 1996.

Among the outcomes of the study is evidence for an expanded definition of aptitude that includes both cognitive aptitude (measured specifically for languages by the MLAT and more generally by cognitive aptitude tests?) and personality factors that predispose a learner to cope with ambiguity and apparent chaos. These become especially important in the relatively unstructured learning setting of communicative teaching approaches. A nexus is emerging of the following characteristics that seem to be related to success in the demanding intensive FSI classroom::

- cognitive aptitude (may include ability to cope with the unfamiliar)
- random (vs. sequential) learning
- orientation to meaning over form
- ability to cope with surprises (linguistic and pedagogical)
- openness to input and tolerance of ambiguity
- ability to sort input, analyze as appropriate, and organize into mental structures.

The last is almost certainly related to the field independence construct in some way; it may be that the MLAT provides a way to measure field independence through verbal activities, in contrast to the usual tests of ability to disembed geometric figures. Such a measure might improve the value of the field independence construct for language learning.

Absence of the above-listed characteristics appears to disadvantage FSI learners, perhaps more than the presence of these variables advantages those learners (Ehrman, 1994a, b, 1995, 1996).

There seems to be a kind of aptitude-personality nexus that consists of cognitive flexibility, tolerance of ambiguity (including ability to impose structure on input), and ability to make use of random access strategies.

The MLAT is the most powerful of the predictive variables used, even in programs that are very different from those in vogue when it was designed. It may be that the ability to manage unfamiliar and contradictory input leads both to success in communicative classrooms and to high scores on the MLAT. The MLAT may gain its relative power because it requires the examinee to cope with the unfamiliar on tasks that at least partially simulate language learning tasks, whereas personality inventories are asking about general life preferences, and strategy inventories do not address how the strategies are used but only whether the student is aware of using them. "Faking good" is nearly impossible on the MLAT, and malingering is vanishingly rare at FSI.

Although the MLAT provides strong information about classroom language learning ability, it is supplemented by personality variables. The significant correlations between the MLAT and the personality measures, though not strong (between .21 and .33), are consistent across personality questionnaire and MLAT subscales (Ehrman 1993, 1994a, b, 1995). In all cases, MLAT scores are linked with variables that suggest tolerance for ambiguity.[10]

The links between the MLAT and personality variables suggest a role for the disposition to use one's cognitive resources in ways that go beneath the surface and that establish elaborated knowledge structures. Those who are open to new material, can tolerate contradictions, establish hypotheses to be tested, focus on meaning, and find ways to link the new with previous knowledge structures seem to have an advantage in managing the complex demands of language and culture learning. The weakest students appear to be overwhelmed by the chaos they encounter; the strongest meet it head on, may even embrace it to a degree.

As of now, the answer to the question "is the MLAT passé?" is: probably not, though it has much the same limitations as a sole *predictor* of learning success that it has always had. It is pretty good, especially if viewed as an indicator of learning dispositions that will affect classroom performance, but it probably should not be more than one tool in a toolkit. Scatter analysis of the part scores is a promising use for placement, counseling, and remediation, particularly in the hands of an evaluator who treats the scores as signposts to interpretations to be tested, not as absolute predictors.

Limitations of this study: The greatest limitation of this study, like all those from FSI, is the question of generalizability. Use of a sample drawn from a high-end, preselected population in itself restricts range, affects distributions, and strongly indicates the need for replication with samples more typical of what the usual reader of this publication works with. For the MLAT,

---

[10] A very recent study also shows a correlation of the MLAT with self-report of 'field sensitivity' (Index, r=.58, Part II .61, Part III .46, all at a p level of 0001). Field sensitivity, discussed at greater length in Ehrman, 1996, in press, and Ehrman & Leaver, 1997, can be defined as preference for working with new material in context, in stories or articles or at least sentences. Field sensitive learners often pick up new words, ideas, etc. peripherally, without planning in advance; they can be described as using of a floodlight to learn in contrast to field independence, which uses a spotlight.

unlike any of the other instruments in the larger study, the use of the instrument itself to help preselect the sample severely limits both the statistical normality of the sample and our ability to make inferences from the findings.

The impossibility of establishing a truly normal distribution of MLAT scores in this sample also means that the statistical tests that assume normal distributions and similar sample sizes are used in unconventional ways. The number of tests conducted increases the chance of type I errors (false positives), though the consistency of findings over a number of variables may reduce the likelihood of such error. For these reasons, the findings reported here must be considered suggestive, not conclusive.

The qualitative investigation has been undertaken on an ad hoc basis and therefore consists for now of working hypotheses about the meanings of high and low points in MLAT part-score profiles. It has yet to be investigated more systematically at a level beyond individual cases.

Next Steps: There is much more to look at in these data, in the course of trying to find out what the MLAT is good for and what its limitations are. Among these are to seek normally distributed samples on which to replicate this study, begin multiple regression and discriminant analysis to see if MLAT is a better predictor in combination with other variables; and to find out what happened with subjects who return from overseas and are tested--did they improve, get worse, stay the same? On the qualitative front, continued investigation can seek to confirm the working hypotheses described above in the section on student counseling and systematize them for use by people other than researchers, so that the MLAT part scores can provide useful information about specific learning strengths and difficulties that can be used in curriculum design and interventions with individual students. Eventually, a quantitative study of the part-score profiles should be designed and undertaken.

References

Adams, M. (1984). The Modern Language Aptitude Test (Percentile Ranks). Unpublished FSI document. Arlington, VA: Foreign Service Institute.

Anastasi, A. (1988). *Psychological testing*. 6th ed. New York: Macmillan.

Campbell, C. (1987). Survey of Attitudes Specific to the Foreign Language Classroom. Unpublished manuscript.

Carroll, J. (1962). The prediction of success in intensive foreign language training. In Glaser, R. (Ed.), *Training research and education*. Pittsburgh, PA: Univ. of Pittsburgh.

Carroll, J. (1990). Cognitive abilities and foreign language aptitude: Then and now. In Parry, T. & C. W. Stansfield (Eds.), *Language aptitude reconsidered*. (pp. 11-29). Englewood Cliffs, NJ: Prentice Hall.

Carroll, J. & Sapon, S.M. (1959). *Modern Language Aptitude Test*. New York: Psychological Corporation.

Clark, J. (1986). *A study of the comparability of speaking proficiency across three government language training agencies*. Washington, DC: Center for Applied Linguistics.

Demuth, K.A., & Smith, N.B. (1987). The foreign language requirement: An alternative program. *Foreign Language Annals, 20*, 67-77.

Ehrman, M.E. (1993). Ego boundaries revisited: Toward a model of personality and learning. In J.E. Alatis (Ed.). *Strategic interaction and language acquisition: Theory, practice, and research*. Washington, DC: Georgetown University Press.

Ehrman, M.E. (1994a). The Type Differentiation Indicator and adult language learning success. *Journal of Psychological Type, 30*, 10 - 29.

Ehrman, M.E. (1994b). Weakest and strongest learners in intensive language training: A study of extremes. In C. Klee (Ed.) *Faces in a crowd: Individual learners in multisection programs*. Boston MA: Heinle & Heinle.

Ehrman, M.E. (1995). Correlations between MLAT scales and preferred student learning activities. Unpublished data.

Ehrman, M.E. (1995). Personality, language learning aptitude, and program structure. Alatis, J. (Ed.), Linguistics and the education of second language teachers: *Ethnolinguistic, psycholinguistic, and sociolinguistic aspects* (pp. 328-345). Washington DC: Georgetown University Press.

Ehrman, M.E. (1996). *Understanding second language learning difficulties*. Thousand Oaks, CA: Sage Publications.

Ehrman, M.E. (in press). Field independence and field sensitivity. In Reid, J. (Ed.), *Centering on the learner*. Englewood Cliffs, NJ: Prentice Hall Regents.

Ehrman, M.E. & Jackson, F.H. (1992). Classroom activities survey. Unpublished manuscript.

Ehrman, M.E. & Oxford, R.L. (1995). Cognition plus: Correlates of language learning success. *Modern Language Journal* (1), 67-89.

Ehrman, M.E., & Leaver, B. L. (1997). Sorting our global and analytic functions in second language learning. American Association for Applied Linguistics annual meeting, Orlando, FL, March 1997.

Hart-Gonzalez, L.H. & Ehrman, M.E. (1992). Study Activities Questionnaire. Unpublished manuscript.

Hartmann, E. (1991). *Boundaries in the mind: A new psychology of personality*. New York: Basic Books.

Oxford, R.L., & Ehrman, M.E. (1995). Adults' language learning strategies in an intensive foreign language program in the United States. *System, 23*, 359-386.

Jackson, F.H. (1994). Language aptitude. Draft talking points prepared for FSI roundtable on the Modern Language Aptitude Test. Annual meeting of the Language Testing Research Colloquium, Washington, DC.

Keefe, J.W. & Monk, J.S., (with Letteri, C.A., Languis, M., & Dunn, R). (1989). *Learning Style Profile*. Reston, VA: National Association of Secondary School Principals.

Lefrancois, J. & Sibiga, T. C. (May, 1986). Use of the Modern Language Aptitude Test (MLAT) as a diagnostic tool. Unpublished paper. Source unknown.

Lett, J.A., & O'Mara, F.E. (1990). Predictors of success in an intensive foreign language learning context: Correlates of language learning at the Defense Language Institute Foreign Language Center. . In Parry, T. & C. W. Stansfield (Eds.), *Language aptitude reconsidered*. (pp. 222-260.) Englewood Cliffs, NJ: Prentice Hall.

Lett, J.A., & Thain, J. (1994). The Defense Language Aptitude Battery: What is it and how well does it work? Paper delivered at the Language Aptitude Invitational Symposium, Arlington VA.

Myers, I.B., & McCaulley, M.H. (1985). *Manual: A guide to the development and use of the Myers-Briggs Type Indicator*. Palo Alto, CA: Consulting Psychologists Press.

Norusis, M. J. (1992). *SPSS for Windows 5.0.1*. Chicago: SPSS Inc.

Parry, T.S. & Child, J.R (1990). Preliminary investigation of the relationship between VORD, MLAT, and language proficiency. . In Parry, T. & C. W. Stansfield (Eds.), *Language aptitude reconsidered*. (pp. 30-66.) Englewood Cliffs, NJ: Prentice Hall.

Parry, T., & Stansfield, C.W. (1990). Introduction. In Parry, T. & C. W. Stansfield (Eds.), *Language aptitude reconsidered*. (pp. 1-10.) Englewood Cliffs, NJ: Prentice Hall.

Parry, T., & Stansfield, C. W. (Eds.) (1990). *Language aptitude reconsidered*. Englewood Cliffs, NJ: Prentice Hall.

Petersen, C.R., & Al-Haik, A.R.. (1976). The development of the Defense Language Aptitude Battery (DLAB). *Educational and psychological measurement, 6*: 369-380.

Pimsleur, P. (1966). *The Pimsleur language aptitude battery*. New York: Harcourt, Brace, Jovanovich.

Pimsleur, P. (1968). Aptitude testing. *Language Learning*. Special Issue Number 3, August 1968, 73-78.

Sasaki, M. (1993). Relationships among second language proficiency, foreign language aptitude, and intelligence: A structural equation modeling approach. *Language Learning 43*, 313-344.

Saunders, D. (1989). *Type Differentiation Indicator Manual: A scoring system for Form J of the Myers-Briggs Type Indicator*. Palo Alto, CA: Consulting Psychologists Press.

Skehan, P. (1989). *Individual differences in second language learning*. London: Edward Arnold.

Skehan, P. (1991). Individual differences in second language learning. *Studies in Second Language Acquisition 13(2)*, 275-278..

Sparks, R. L., Ganschow, L . and Patton, J. (1995). Prediction of performance in first-year foreign language courses: Connections between native language and foreign language learning. *Journal of Educational Psychology, 87*: 638-655.

Spolsky, B. (1995). *Measured words: The development of objective language testing*. New York: Oxford.

Thain, J. (1992). DLAB II prototype development: Status report and CY92 plan. Technical report. Monterey, CA: Defense Language Institute.

Wesche, M. B. (1981). Language aptitude measures in streaming, matching students with methods, and diagnosis of learning problems. In K.C. Diller (Ed.), *Individual differences and universals in language learning aptitude*. Rowley, MA: Newbury House.

Wesche, M. B., Edwards, H., & Wells, W. (1982). Foreign language aptitude and intelligence. *Applied Psycholinguistics, 3*, 127-40.

Wesche, M., Edwards, H., & Wells, W. (1982). Foreign language aptitude and intelligence. *Applied Psycholinguistics*, 127-140.

Wilds, C. P. (1965). MLAT Index Scores: Foreign Service Institute Curve - 1965. Unpublished table. Arlington, VA: Foreign Service Institute.

Appendix A: Conversion Table for MLAT Raw Total and Index Scores

Raw Total	Index	Raw Total	Index	Raw Total	Index
0-9	15	67-68	37	125-127	59
10-12	16	69-71	38	128-129	60
13-15	17	72-74	39	130-132	61
16-18	18	75-76	40	133-135	62
19-21	19	77-79	41	136-137	63
22-23	20	80-82	42	138-140	64
24-26	21	83-84	43	141-143	65
27-29	22	85-87	44	144-145	66
30-31	23	88-90	45	146-148	67
32-34	24	91-92	46	149-150	68
35-37	25	93-95	47	151-153	69
38-39	26	96-97	48	154-156	70
40-42	27	98-100	49	157-158	71
43-44	28	101-103	50	159-161	72
45-47	29	104-105	51	162-164	73
48-50	30	106-108	52	165-166	74
51-52	31	109-111	53	167-169	75
53-55	32	112-113	54	170-172	76
56-58	33	114-116	55	173-174	77
59-60	34	117-119	56	175-177	78
61-63	35	120-121	57	178-180	79
65-66	36	122-124	58	181-180	80

From Wilds (1965)

APTITUDE TESTS: CONCEPTION AND DESIGN

James R. Child
Department of Defense

Language Aptitude Testing: Language Learners and Language Applications

Language Learning: Available populations

Language learners come in the main from two quarters within US government agencies: the onboard cadre and prospective hires. The problems and promise of each are considered below, with English understood as the first language of most prospective examinees.

Onboard working linguists

Of the onboard force those persons already productively engaged in second language work are, if they can be spared, the best bets for cross-training. Of these persons it can be safely said that certain ones are better candidates for the "most difficult" languages, while others may be retrained in those third languages which are at "medium" distance from the ones they are currently working. In either case, the productive linguist may have credentials equal to those that present "aptitude" measures (MLAT, DLAB, et al) can confer.

Prospective linguists

For present purposes these are by and large new hires who have excellent academic records in western European languages, but are scheduled for retraining in "middle" or "remote" languages. They are supplemented on occasion by a small number of onboard non-linguists who for some reason need to have the "elements" of one such language. Such persons should be tested for aptitude so that managers have at least some idea of the odds of success.

Language uses: skills and levels

Of the four skills described in the present version of the ILR guidelines, only three come into play with any frequency: speaking, non-interactive listening and reading. As for levels, these are best determined by the kinds of texts the language learners

can be expected to comprehend and/or produce in the three mentioned skills. For most purposes full-range level 2 attainments in one or more of the three are absolute minimums.

While there is value in using aptitude measures with persons scheduled o study difficult languages there seem to be no tests specifically designed to predict success beyond level 2. It is at level 3 (or perhaps 2+) that major cultural differences between languages begin to cause difficulties for learners in various semantic areas. The fact is, however, that most prospective learners need to develop skills within the range of level 3 and (ideally at 3+ or 4) if the Government is to get its money's worth from training. Whether aptitude measures already exist in some form, or can be developed to address this question is uncertain. Other kinds of cultural-sensitivity models are available and may be the best vehicles for the purpose.

Language channels: speech vs. writing systems

Channels for present purposes are those means by which language can be delivered, i.e., through speech or writing. The skills required for the respective channels are tied in large measure to the differing social dynamics characterizing each.

Speech

Speech can be viewed for any language as the starting point of the whole of communication. It is usually in the form of an exchange in which initiation, response and rejoinder occur without elaborate planning. Viewed in this way, language use becomes an arena of action in which rapid shifts from production to reception and back are the norm for communication and a kind of standard for language learners. That is, the language learners for whom the "speech" channel will be central to their experience should be psychologically as well as linguistically prepared for conversational give-and-take. Naturally, a complex of skill entailing speech production and aural comprehension (in ILR terms, Speaking and Interactive Listening) will enter into whatever kind of aptitude test is developed.

It is not true, however, that speech invariably requires both production and reception skills. There are situations in everyday life where speaking does not assume an oral response: persons listening to radio and television broadcasts, or to lectures in an auditorium, are in on position to respond immediately to what they are hearing (although there may be opportunities later to call into the station with comments or put questions to a speaker at the end of a presentation). Even in these cases,

though, processing of speech in a delayed response mode still requires the skills associated with the conversion of sound to meaning, supported, to be sure, by tape replay when this is possible or feasible. Memory is in any case an essential in capturing meaning from the flow of sound through time. In situations where language performance rather than general proficiency is demanded aptitude testing of a highly specialized nature may be required. Specialized aptitude instruments of this sort are rare commodities.

Writing systems

Writing systems are of relatively recent origin and naturally derivative, although they can rapidly take on lives to some degree independent of the spoken language on which they are based. Thus, in many instances alphabets and scripts may not track with the phonology of the spoken language because they have been frozen for decades or centuries while the spoken language evolves rather rapidly. As for "character sets" associated with languages such as Chinese or ancient Egyptian, there is little or no phonic/graphic linkage. However, one factor characterizes written as opposed to spoken texts: space rather than time permits easier accessibility to processing information. Current aptitude models by and large build on that fact.

Reading/speaking crossover

The complex relationships of the two skills has important implications for aptitude testing. Language strings in written form can be described as "flowing through space" rather than time, which allows the reader easier access to preceding and following material noted above than is the case with speech. Offsetting the advantage thus conferred, however, are problems inherent in the sometimes tenuous relation between the phonology of the spoken and written systems alluded to above. Questions of grammar and syntax enter the picture, too, but they are not simply concerned with delivery channels: language in spoken form may actually have its source in a written text, and conversational materials is sometimes reduced to written form, to be read later. The challenges to comprehension of mixed modes of delivery thus entail at some point the requirement to surmount the difficulties of the flow of spoken language in internalizing, and often making a record of the processed material in answer to whatever style and register the text is couched.

Clearly an aptitude test which contains word-through clause-level material at most will not get at anything more than the phono-morphological structure (i.e. the

"canonic forms") of an artificial language in which the rules of speech are detailed in the exercise. In the case of a natural language, an extended period of familiarization with that language would be needed to test even the simplest of utterances; this is a luxury normally not available for aptitude measures. It might be possible to devise an aptitude model in which the grammar and lexicon of a given written language are detailed but which requires the examinee to recover somewhat variant forms and junctures typical of speech.

Distance between languages

The retraining of linguists in other languages was raised above in connection with the needs of the work force. The critical question here, however, is the distance between whatever language skills the learner has already mastered (including skill levels in the native language) and those skills needed in acquiring a new language or languages. For example, a linguist either newly hired or on board for some time has a good reading knowledge of Chinese. However, local need requires reading skill in Japanese; the character set the learner already controls from Chinese can be most serviceable for the latter language. Managers can use information of this kind in planning retraining and can likewise make use of the studies on language distance now or prospectively employed by the Government language schools. Time does not permit extended commentary on distances between languages but a few observations may suffice, based on an approach to language "difficulty" under consideration for government-wide use.

The system currently in place for dealing with language "difficulty"—in practical terms, the problems native speakers of English have in mastering other languages—provides for four categories, from "easiest" to "hardest"— without greater detail. There is now available a matrix which attempts to identify just what is difficult for an American student of a second (or third) language. The matrix lists in a vertical column critical language elements, i.e., phonic and graphic systems (Block A); grammatical systems (Block B); and semantic/cultural systems (Block C). The horizontal axis specifies presumed distances from English: 1–Near; 2–Middle; 3–Remote. The resulting nine cells contain explanatory material relevant to each cell as it applies (tentatively) to one aspect of some 115 languages. Thus, Japanese may be summed up alphanumerically as A3, B3, C3: a language whose systems of writing, grammar and conceptualization are all (relatively) remote from English. Chinese, on the other hand, while sharing with Japanese the complexities of the writing system, is not as remote in regard to its grammatical and (possibly) its semantic system. Thus, it could be represented as A3, B2, C2.

This study is in its preliminary stages, and many changes will be forthcoming. As a beginning, though, it deserves consideration in the framework of language aptitude theorizing. A copy of this matrix is available on request.

Summary

From the above it should be clear that there are complex relationships between the backgrounds and attainments of prospective second (or third) language learners and the frames of reference in which those learners may be expected to operate. Thus the notion of "language aptitude" is to be considered in light of the level of linguistic and cultural skill required of or desired by the learner; the channel (speech, writing) in which the learner is more comfortable; and his or her need on occasion to deal with both channels in (roughly) the same time frame. Clearly a variety of instruments is needed some of which will demand much more testing time than those currently in use and will discomfit supervisory and other personnel who prefer short and snappy tests.

APTITUDE FROM AN INFORMATION PROCESSING PERSPECTIVE

Barry McLaughlin
University of California, Santa Cruz

For some years now I have been wrestling with the question of aptitude from within an information processing perspective. In this paper I will briefly outline the approach that I take, examine how aptitude is conceptualized in this framework, and discuss one possible component of second language aptitude, working memory.

Information Processing

Because human learners are limited in their information-processing abilities, only so much attention can be given to the various components of complex tasks at one time. In order to function effectively humans develop ways of organizing information. Some tasks require more attention; others that have been well practiced require less. The development of any complex cognitive skill involves building up a set of well-learned, efficient procedures so that more attention-demanding processes are freed up for new tasks. In this way limited resources can be spread to cover a wide range of task demands.

In this framework, learning is a cognitive process because it is thought to involve internal representations that regulate and guide performance. In the case of language learning, these representations are based on the language system and include procedures for selecting appropriate vocabulary, grammatical rules, and pragmatic conventions governing language use. As performance improves (becomes more automatic), there is constant restructuring as learners simplify, unify, and gain increasing control over their internal representations (Karmiloff-Smith 1986). These two notions—automatization and restructuring—are central to the information processing approach.

The Routinization of Skills

Several researchers (Hasher and Zacks 1979, Posner and Snyder 1975, Schneider and Shiffrin 1977, Shiffrin and Schneider 1977) have conceived of the differences in the processing capacity necessary for various mental operations in a dichotomous way: either a task requires a relatively large amount of processing capacity, or it proceeds automatically and demands little processing energy. Furthermore, a task that once taxed processing capacity may become, through practice, so automatic that it demands relatively little processing energy.

Automatic processing involves the activation of certain nodes in memory each time the appropriate inputs are present. This activation is a learned response that has been built up through the consistent mapping of the same input to the same pattern of activation over many trials. Because an automatic process utilizes a relatively permanent set of associative connections in long-term storage, most automatic processes require an appreciable amount of training to develop fully. Once learned, however, automatic processes occur rapidly and are difficult to suppress or alter.

The second mode of information processing, controlled processing, is not a learned response, but instead a temporary activation of nodes in a sequence. This activation is under the attentional control of the subject and, because attention is required, only one such sequence can normally be controlled at a time without interference. Controlled processes are thus tightly capacity-limited, and require more time for their activation. But controlled processes have the advantage of being relatively easy to set up, alter, and apply to novel situations. The clearest example of this distinction that I can think of is writing with one's right and left hand. Assuming that you are a right-handed person, writing with that hand is automatic, but writing with the left hand requires controlled processing.

Consider the following report of a schizophrenic patient.

> I'm not sure of my own movements any more.... I found recently that I was thinking of myself doing things before I would do them. If I'm going to sit down for example, I've got to think of myself and almost see myself sitting down before I do it. It's the same with other things like washing, eating, and even dressing—things that I have done at one time without even bothering or thinking about at all ...I take more time to do things because I am always conscious of what I am doing. If I could just stop noticing what I am doing.... I have to do everything step by step now, nothing is automatic. Everything has to be considered (from McGhie 1969).

Of course, this is a very dysfunctional situation. If we had to think through ordinary activities before we did them, we would not be able to manage our lives very well.

What we see in this patient is a breakdown in the automaticity that is so important for normal functioning. We perform numerous complex tasks in our daily lives automatically, without thinking about them. But this was not always the case; we had to learn to perform the operations involved in these complex skills by focusing attention on them.

Learning to drive using a clutch or attempting to master the backhand in tennis are tasks that require a great deal of attention—or what I am referring to here as "controlled processing." After one has practiced the task, components of these skills become automatic, and controlled processing is required only in unusual cases. When you have been driving for many years, you can carry on a conversation as long as no emergencies arise; but if you have to drive on a very icy road, controlled processing is called into play and it is difficult to keep a conversation going.

With enough practice, it is possible for people to carry out quite amazing feats. In one experiment, after extended practice, subjects were able to read a story aloud while writing down another story from dictation (Solomons and Stein, 1896, cited in Howard 1983). In this case, presumably, reading had become so automatic that the subjects could devote attention to the other task. Note that from an information-processing perspective, the same principles apply to complex skills such as reading, writing, or learning a second language as apply in the case of motor skills such as driving, typing, or playing tennis.

In short, within this framework, complex cognitive skills are learned and routinized (i.e., become automatic) through the initial use of controlled processes. Controlled processing requires attention and takes time, but through practice sub-skills become automatic and controlled processes are free to be allocated to higher levels of processing. Thus controlled processing can be said to lay down the "stepping stones" for automatic processing as the learner moves to more and more difficult levels (Shiffrin and Schneider 1977).

In this conceptualization, complex tasks are characterized by a hierarchical structure. That is, such tasks consist of sub-tasks and their components. The execution of one part of the task requires the completion of various smaller components. As Levelt (1978) noted, carrying on a conversation is an example of a hierarchical task structure. The first-order goal is to express a particular intention. To do this, the speaker must decide on a topic and select a certain syntactic schema. In turn, the realization of this schema requires sub-activities, such as formulating a series of phrases to express different aspects of the intention. But to utter the phrases there is the need for lexical retrieval, the activation of articulatory patterns, utilization of appropriate syntactic rules, etc. Each of these component skills needs to be executed before the higher-order goal can be realized, although there may be some parallel processing in real time.

Note the importance, in this framework, of practice. The development of any complex cognitive skill is thought to require building up a set of well-learned, automatic procedures so that controlled processes are freed for new learning. From a practical standpoint, the necessary component is overlearning. A skill must be practiced again and again and again, until no attention is required for its performance. *Repetitio est mater studiorum*—practice, repetition, time on task—these seemed to be the critical variables for successful acquisition of complex skills, including complex cognitive skills such as second-language learning.

This conceptualization, however, leaves something out of the picture, and runs contrary to the experience of researchers in the second-language field. As Patsy Lightbown wrote in a review paper:

> *Practice does not make perfect.* Even though there are acquisition sequences, acquisition is not simply linear or cumulative, and having practiced a particular form or pattern does not mean that the form or pattern is permanently established. Learners appear to forget forms and structures which they had seemed previously to master and which they had extensively practiced. (Some researchers have referred to 'U-shaped development.')

She went on to discuss some of her own research:

> Learners were—for months at a time—presented with one or a small number of forms to learn and practice, and they learned them in absence of related contrasting forms. When they did encounter new forms, it was not a matter of simply adding them on. Instead the new forms seemed to cause a restructuring of the whole system (Lightbown, 1985, p.177).

Restructuring and Reading

These comments made sense, and helped clarify some puzzling data from a study of second-language reading (McLeod and McLaughlin 1986). The data came from an analysis of errors that speakers of differing degrees of proficiency in English made when reading aloud. We found that the errors that beginning ESL students made were primarily nonmeaningful, which was seen to be due to these students focusing on the graphic aspects of the text. That is, they would make errors like "She shook the piggy bank and out came some many" (for 'money'), whereas native speakers were more likely to make meaningful errors, such as "She shook the piggy bank and out came some dimes." It was expected that the proportion of meaningful errors for advanced ESL students would fall somewhere between what was found for beginning ESL students and native speakers. But instead, it was found that advanced ESL students, who had a much superior grasp of the syntactic and semantic constraints of English (as shown by their performance on a cloze test), made as many nonmeaningful errors as the beginning students.

Research on reading indicates that beginning readers who have mastered the mechanical aspects of reading continue to process the text word by word, not using contextual semantic relations and syntactic information to comprehend meaning (Cromer 1970). What was surprising to us was that more advanced second language learners in our study were apparently doing the same thing. Their errors showed that they were not utilizing semantic and syntactic cues as well as they could have. They were not approaching the task as "a psycholinguistic guessing game," in which graphic cues were used to make predictions about what the printed text means—even though the evidence from the cloze test suggested that they were quite capable of making such predictions. Their increasing syntactic and semantic competence enabled them to make nearly twice as many accurate predictions as the beginners on the cloze test. Yet they had not applied this competence to their reading behavior.

This suggests a process of restructuring had not yet occurred. What seemed to be happening was that the advanced subjects were using old strategies aimed at decoding in a situation where their competencies would have allowed them to apply new strategies directed at meaning. Their performance on the cloze test indicated that they had the skills needed for "going for meaning." Presumably they read this way in their first language. But they had not yet made the shift (restructured) n their second language. In this language, they did not make strategic use of the semantic and syntactic knowledge at their disposal. Indeed, other researchers obtained very similar results in second-language reading (Clark 1979).

The Restructuring Concept

The concept of restructuring can be traced in the psychological literature to the developmental psychologist, Jean Piaget. The Piagetian structuralist approach maintains that cognitive development is an outcome of underlying structural changes in the cognitive system. Just what constitutes structural change has been a topic of some debate (see Globerson 1986; Karmiloff-Smith 1986). Suffice it to say that there appears to be agreement that not just any change constitutes restructuring. Restructuring is characterized by discontinuous, or qualitative, change

as the child moves from stage to stage in development. Each new stage constitutes a new internal organization and not merely the addition of new structural elements.

Recent concern with restructuring in developmental psychology reflects a new emphasis on the dynamics of change and a reaction to what had became known as the "snapshot problem." That is, developmental psychologists became concerned that their knowledge of cognitive growth consisted of a series of "snapshots" of the child's abilities at various points in development, but that they knew little about how the child progressed from snapshot to snapshot. The analogy in the field of second-language research is the concern—expressed by a number of authors (e.g., Hatch 1978; Huebner, 1983; Long and Sato 1984)--that there is more known about linguistic products, but little known of the dynamics of psycholinguistic processes.

From an information processing perspective, restructuring can be seen as a process in which the components of a task are coordinated, integrated, or reorganized into new units, thereby allowing the procedure involving old components to be replaced by a more efficient procedure involving new components (Cheng 1985). To study restructuring is to focus on the mechanisms of transition that are called into play as the learner modifies internalized, cognitive representations.

In short, learning inevitably goes beyond mere automaticity. There is a constant modification of organizational structures. Rumelhart and Norman (1978) identified restructuring as a process that occurs "when new structures are devised for interpreting new information and imposing a new organization on that already stored" (p. 39). They contrasted this process of learning with (a) accretion, whereby information is incremented by a new piece of data or a new set of facts, and (b) tuning, whereby there is a change in the categories used for interpreting new information. In tuning, categories, or schemata, are modified; in restructuring, new structures are added that allow for new interpretation of facts.

Rumelhart and Norman argued that learning is not a unitary process, but that there are different kinds of learning, one of which is restructuring. Whereas some learning is thought to occur continuously by accretion, as is true of the development of automaticity through practice, other learning is thought to occur in a discontinuous fashion, by restructuring. This discontinuity accounts for the second-language learner's perceptions of sudden moments of insight or "clicks of comprehension." At such moments, presumably, the learner can be said to understand the material in a new way, to be looking at it differently. Often learners report that this experience is followed by rapid progress, as old linguistic information and skills are fit into this new way of understanding. As Kolers and Roediger (1984) put it, learning involves a reassembly and refinement of procedures of the mind.

Second-Language Learning As a Complex Cognitive Skill

Applying these notions more specifically to second-language learning, one can say that from an information-processing perspective, second-language learning, like any other complex cognitive skill, involves the gradual integration of sub-skills, as controlled processes initially predominate and later become automatic. Thus the initial stages of learning involve the slow development of skills and the gradual elimination of errors as the learner attempts to automatize aspects of

performance. In later phases, there is continual restructuring as learners shift their internal representations. Although both processes occur throughout the learning of any complex cognitive skill, gains in automaticity are thought to be more characteristic of early stages of learning and restructuring of later stages.

For the most part, second-language researchers have been more concerned with the development of automaticity than with restructuring, though there has been some recognition of the role restructuring plays in second-language acquisition. A number of authors have commented on discontinuities in the second-language learning process (e.g., Pike 1960, Selinker 1972). Lightbown (1985) pointed out that second-language acquisition is not simply linear and cumulative, but is characterized by backsliding and loss of forms that seemingly were mastered.

Restructuring provides an explanation for examples of U-shape developmental functions in language learning, where performance declines as more complex internal representations replace less complex ones, and increases again as skill becomes expertise. There are many examples of such U-shaped functions in the literature on first- and second-language learning (see McLaughlin 1990). One example is a common strategy adopted by young second-language learners (and, perhaps by more older second-language learners than we realize) to memorize formulas (Hakuta, 1976, Wong Fillmore 1976). Some children are capable of amazing feats of imitation, producing multi-word utterances, which, it turns out, they understand only vaguely. Such unanalyzed chunks appear to show evidence of a sophisticated knowledge of the lexicon and syntax, but it has become clear that such holistic learning is a communicative strategy that second-language learners use to generate input from native speakers (Wong Fillmore 1976).

Subsequently, such formulas are gradually "unpacked" and used as the basis for more productive speech. At this stage, the learner's speech is simpler but more differentiated syntactically. Whereas utterances were as long as six or seven words in the initial stage, they are now much shorter. The learner has at this point adopted a new strategy, one of rule analysis and consolidation.

Expert Systems

Now I would like to turn to the question of what makes a successful language learner. When I ask my students this question, they inevitably answer that you have to have an ear for languages. Some of them say they do not have such an "ear," and cannot learn second languages. As researchers and practitioners, we know that there is not much evidence for a "language ear." There has been some work on the relationship between learning a second language and musicality, but correlations are modest or nonexistent. This is generally true of research directed at the personality characteristics of "the good language learner." In this tradition, researchers studied traits of good learners—musicality, intelligence, extraversion, empathy, and the like. Other researchers examined self-esteem (Heyde, 1977), tolerance of ambiguity (Naiman, Frohlich, Stern, and Todesco, 1978) or the role of motivational and attitudinal variables (Gardner and Lambert, 1972; Nelson and Jakobovits, 1970).

I (Nation and McLaughlin 1986) have argued that three problems beset these efforts. First, there is the problem of the difficulty of obtaining valid and independent measures of personality traits and motivational variables (Oller, 1981). There is considerable shared variance between many of these variables and it is difficult to tease out effects due to each separately. Second, there is the issue of trait by instruction interactions (McLaughlin, 1980). In the real world, it may be that some instructional methods work better than others for individuals with certain personality traits. The good language learner in one context may not be a good language learner in another. Finally, there is the question of causal direction. There may in fact be instances where the direction of causality is from learning to personality factors rather than the other way around. Some evidence for this notion comes from studies on attitudes and language learning in children (Hermann, 1980; Strong, 1984), which suggest that acquiring skill in a languages influences attitudes toward acquisition.

Because of the problems inherent in an approach that looks at person factors, my colleagues and I have taken another tack—one that focuses on process. Specifically, we suggest that "expert" language learners use different information-processing strategies and techniques than do more "novice" learners. This appears to be true in other domains. For example, Chase and Simon (1973) replicated de Groot's(1965) finding that Master chess players reconstructed with greater than 90 percent accuracy midgame boards they had seen for only five seconds. They observed that Master players recalled clusters that formed attack or defense configurations, whereas beginners lacked the skill to form such abstract representations. Strategy differences were also reported by Adelson (1981), who found that expert computer programmers used abstract, conceptually based representations when attempting to recall programming material, whereas novices used more concrete representations. Differences between experts and novices have also been found in research on learning mechanisms in physics (Chi, Glaser, and Rees, 1981), arithmetic (Brown and Burton, 1978), algebra (Lewis, 81), and geometry (Anderson, Greeno, Kline, and Neves, 1981). For the most part, these studies show that experts restructure the elements of a learning task into abstract schemata that are not available to novices, who focus principally on the surface elements of a task. Thus experts replace complex sub-elements with single schemata that allow more abstract processing.

In the realm of language learning, we argued that experts are those individuals who have learned a number of languages. There is considerable anecdotal evidence that once a person has learned a few languages, subsequent language learning is greatly facilitated. Presumably, there is some positive transfer that results from the process of language learning and carries over to the learning of a new language. Unfortunately, there is very little experimental evidence for such a positive transfer hypothesis. Hence, we have conducted a number of studies using miniature linguistic systems to ascertain what makes more experienced language learners different from novices.

Nation and McLaughlin (1986) carried out an experiment in which we contrasted information processing in multilingual, bilingual, and monolingual subjects learning a miniature linguistic system. We wanted to see how "expert" language learners (multilingual subjects) compared in their performance with more "novice" language learners. Subjects were asked to learn a finite-state Markov grammar under conditions in which they were merely exposed to the system without

instructions to learn it (Implicit learning) or under conditions in which they were told that the system was rule-based and they should learn the rules (Explicit learning).

Multilingual subjects were found to learn the grammar significantly better than bilingual or monolingual groups when the instructions called for "Implicit" learning, but not when the instructions called for "Explicit" learning. We argued on the basis of the subsequent analyses that the superior performance of the multilingual subjects on the Implicit-learning task was the result of better automated letter- and pattern-recognition skills.

In general, it may be that individuals with more language-learning experience build up certain basic skills that transfer to new language-learning situations. These skills might include automated auditory recognition skills, pattern recognition skills, word-decoding skills, and superior auditory memory. Because these sub-skills of the task have become relatively automatic in multilingual subjects, attention is freed up to be devoted to the recognition of rule-governed regularities.

In another experiment from our laboratory (Nayak, Hansen, Krueger, and McLaughlin, 1990), monolingual and multilingual subjects were exposed to a limited subset of permissible strings from an artificial linguistic system. We were interested in whether they could apply generalizations derived from the learned subset to novel strings and if so, what was the nature of these generalizations. Subjects were exposed to "sentences" in a grammar ranging in length from two to five words. Words were CVC trigrams. Above each word abstract forms appeared, which were the referents for that word. Subjects were assigned at random to one of two learning conditions: (a) a memory condition, in which they were told to memorize the strings, or (b) a rule-learning condition, in which they were told to look for underlying rules. Subsequent to the learning phase, subjects were shown abstract forms coupled with CVC words and were asked to decide whether the word was matched with the correct form. Subjects were also asked to decide whether novel strings were acceptable in the linguistic system they had been exposed to.

The results of this study indicated that there were differences in learning, in that subjects in the memory condition did better on the vocabulary task than did subjects in the rule-learning condition, while the reverse was true for decisions about the acceptability of novel strings. However, the general level of performance of multilingual and monolingual subjects did not differ: the "experts" were not better than the "novices" in either the vocabulary or the syntax acceptability task, but there were differences in how the two groups went about the tasks.

To examine these differences we had asked subjects at three points during the learning phase to verbalize for another potential subject exactly what they were doing and what strategies they were using. We coded the verbalizations of all subjects into four categories. The first two referred to strategies that involved the use of **mnemonic** devices, either *visual*—for example:

> First, I was trying to look at the abstract form above the word and just try and remember what they looked like, see if there's some type of correlation....CAV looked like a cave. KOR was similar to a Russian word, I tried to associate the words with the symbols.

or *verbal*—e.g.:

> ... I tried looking at the words themselves and seeing if I could eliminate certain letters, like the first letter of each thing, if I could form a word out of that or I kept just trying different combinations to see if by reading it backwards and forwards and all these different ways, if it would make some sense.

Two other categories reflected the use of **linguistic** strategies, either *structural*:

> This time it seems like I'm inclined to finding places more than... I uh, it seems like I'm splitting them up into nouns and verbs and objects, and if one goes into a place where, say if an object goes into a place where I think a verb should be, I think it shouldn't be there....

or referring specifically to *word order*:

> I still feel like the rectangular one goes at the beginning, and then either the straight line or the zigzag lines comes in second place, and then the CAV or DUP usually are at the end. When they come into the middle, I feel like the sentence isn't in its proper order.

We found that multilingual subjects were more likely to use mnemonic devices than linguistic strategies in the memory condition, but that in the rule-discovery condition, both groups of subjects preferred linguistic strategies to mnemonic devices, although the difference was statistically significant only for the multilingual subjects.

In addition, we found that multilingual subjects used a wider variety of different strategies in the rule-discovery than in the memory condition, and that no such difference existed for the monolingual subjects. This suggests that one difference between more and less experienced language learners relates to flexibility in switching strategies. This is consistent with the research of Nation and McLaughlin (1986), who found that multilingual subjects were able to avoid perseveration errors more than were other subjects in their experiment. Similarly, Ramsey (1980) reported that multilingual subjects demonstrated greater flexibility in "restructuring mental frameworks" than did monolingual subjects.

Thus there is some evidence to suggest that more expert language learners show greater plasticity in restructuring their internal representations of the rules governing linguistic input. This ability to exert flexible control over linguistic representations and to shift strategies may result from "learning to learn," in the sense that experience with a number of languages may make the individual more aware of structural similarities and differences between languages and less constrained by specific learning strategies. More experienced learners may more quickly step up to the metaprocedural level and weigh the strategies and tactics they are using.

Such a conclusion needs to be tempered by noting that in our research the differences between more and less successful language learners were relatively subtle. There are still many questions

remaining—especially the question of what makes it possible for some individuals to be more flexible than others in forming mental representations of a new linguistic system? What is the reason for differences in information-processing strategies? This brings me to the issue of aptitude. Specifically, I would like to talk about a possible components of language aptitude, working memory.

Working Memory and Aptitude

At first glance, introducing the notion of aptitude may appear to lead us back to innate personality traits, but, as I will argue in more detail later, it is not necessary to think of aptitude as a fixed capacity. Instead, I suggest that we conceptualize aptitude as modifiable by previous learning and experience. Novices can become experts with experience. In the case of multilinguals, experience with several languages provides them with strategies and metacognitive skills that generalize to subsequent languages.

The classic work on aptitude, of course, is that of John B. Carroll. In a 1981 paper, Carroll argued that the tasks contained in aptitude tests are similar to the processes described in information-processing accounts of cognitive functioning. He speculated, for example, that individual variation in the ability to recognize grammatical functions and to match functions in different sentence structures may reflect differences in the ability to operate in "executive" working memory and to store and retrieve information from short-term memory. This line of thinking anticipated recent work on expert systems and is quite consistent with the framework I am advocating. In particular, recent work on the concept of "working memory" fits in very well with the way Carroll conceptualized language learning aptitude.

The concept of working memory is relatively new in cognitive psychology, and refers to the immediate memory processes involved in the simultaneous storage and processing of information in real-time. The term dates to Newell (1973) and is distinguished from the more traditional understanding of short-term memory as a passive storage buffer. Working memory is assumed to have processing as well as storage functions; it serves as the site for executing processes and for storing the products of these processes. For example, in processing a second language, the learner must store phonological, syntactic, semantic, and pragmatic information and must use this information in planning and executing utterances. This information can become a part of working memory via a number of routes: it may be perceptually encoded from the input of an immediate interlocutor, it may be sufficiently activated so that it is retrieved from long-term memory; or it can be constructed as speech is planned.

Working memory is assumed to be limited in its capacity. Working memory limits constrain the development of complex cognitive tasks at several stages. Assuming that the mastery of such complex tasks requires the integration of controlled and automatic processing (e.g., LaBerge and Samuels, 1974; Schneider and Shiffrin, 1977; Shiffrin and Schneider, 1977), one would expect that more working-memory capacity is required at the attention-demanding initial phase when controlled processes predominate. Later, when subtasks that once taxed processing capacity become so automatic that they require little processing energy, working-memory load is reduced.

A second limitation of working memory on the acquisition of a complex cognitive skill occurs later as automaticity builds up and memory load is reduced. Anderson (1983) has suggested that, although initial formation of automatic processes reduces working-memory load, subsequent skill improvement actually increases working-memory load. The reason for this is that the size of subtasks (or what Anderson calls "composed productions") increases. Larger subtasks require more conditions to be active in working memory before they can execute. This may be an explanation for the kind of effects that occur as learners impose organization on information that has been acquired.

What I am suggesting is that increased practice can lead to improvement in performance as sub-skills become automated, but it is also possible for increased practice to create conditions for restructuring with attendant decrements in performance as learners reorganize their internal representational framework. In the second case, performance may follow a U-shaped curve, as was discussed earlier, declining as more complex internal representations replace less complex ones, and increasing again as skill becomes expertise. The reason for such U-shaped functions is that integrating large subtasks makes heavy demands on working memory, and hence performance is actually worse in subsequent stages than it is initially.

Recent research on working memory shows links to vocabulary development, speech production, comprehension, and phonological memory (Gathercole and Baddeley, 1993; Harrington, 1992; Service, 1992; Speidel, 1993). Nonetheless, several hypotheses remain to be tested: first, the question of whether working memory in the first and second language are independent or two aspects of the same thing. If relative processing efficiency is independent of specific language development, it is expected that relative working memory capacity in the first language will also be evident in the second. Further, it would be expected that individuals with larger first language working memory capacity will be better, possibly faster learners of the second language. Second, there is the question of the development of second-language working memory across time. Longitudinal studies are needed to provide a profile of how second-language working memory capacity and second-language proficiency co-vary in the course of development. Such research is important to demonstrating a causal link between working memory and second-language learning.

Can Aptitude Be Learned?

Although much research remains to be done on the role of working memory in second-language learning, I suspect that individual differences in language learning aptitude are due in large measure to the joint function of availability of knowledge about the target language and the speed and efficiency of working memory—which affects the extent to which the individual succeeds in generating and altering the cognitive data required at various processing stages. That is, in second-language learning working memory relates to the degree to which individuals can more flexibly and consistently restructure and reconfigure linguistic representations.

I believe that there are strategies that can be taught to increase the efficiency of working-memory processes. Indeed, within an "expert systems" framework, Faerch and Kasper (1983), McGroarty (1989), Oxford (1986), and O'Malley and Chamot (1989) have attempted to specify strategies that good language learners use and to teach them to less expert learners. The ultimate goal of

much of this research has been to expand and refine the repertoire of strategies of poor learners so that they may benefit from strategies used to good effect by "expert" learners. This work on strategy differences that distinguish good from poor language learners (O'Malley and Chamot, 1989; Oxford, 1986; Wenden, 1987) is important for teaching learners to be more efficient information processors. Indeed, experimental research (e.g., Chase and Ericcson, 1982) has shown that working-memory capability can be greatly expanded as a function of relevant knowledge structures and strategies.

In conclusion, I think there is an answer to students who complain that they just do not have any aptitude for languages. My response is that, although some students definitely have an advantage in language learning because of strategies they have developed and their knowledge base, this does not mean that other students cannot develop similar strategies and build up their knowledge base. Of course the goal for researchers and practitioners is to identify the relevant strategies and help students use them to build up their knowledge of the language and skill in using it.

References

Adelson, B. (1981). Problem Solving and the Development of Abstract Categories in Programming Languages. *Cognition, 9*, 422-433.

Anderson, J. (1983). *The Architecture of Cognition*. Cambridge, MA: Harvard University Press.

Anderson, J. R., Greeno, J. G., Kline, P. J., and Neves, D. M. (1981). Acquisition of Problem-Solving Skills. In J. R. Anderson (Ed.), *Cognitive Skills and Their Acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Brown, J. S., and Burton, R. R. (1978). Diagnostic Models for Procedural Bugs in Basic Mathematical Skills. *Cognitive Science, 2*, 155-192.

Carroll, J. B. (1981). Twenty-Five Years of Research on Foreign Language Aptitude. In K. C. Diller (Ed.), *Individual Differences and Universals in Language Learning Aptitude*. Rowley, MA: Newbury House.

Chase, W. C., and Ericcson, K. A. (1982). Skill and Working Memory. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation*, 16, New York: Academic Press.

Chase, W. C., and Simon, H. A. (1973). Perception in Chess. *Cognitive Psychology, 4*, 55-81.

Cheng, P. W. (1985). Restructuring Versus Automaticity: Alternative Accounts of Skill Acquisition. *Psychological Review, 92*, 214-223.

Chi, M., Glaser, R., and Rees, E. (1981). Expertise in Problem Solving. In *Advances in the Psychology of Human Intelligence*. (Vol. 1). Hillsdale, NJ: Lawrence Erlbaum Associates.

Clark, M.A. (1979). Reading in Spanish and English: Evidence from Adult ESL Students. *Language Learning, 29*, 121-47.

Cromer, W. (1970). The Difference Model: A New Explanation for Some Reading Difficulties. *Journal of Educational Psychology, 61*, 471-483.

de Groot, A. D. (1965). *Thought and Choice in Chess*. Paris: Mouton.

Faerch, C. and G. Kasper (Eds.) (1983). *Strategies in Interlanguage Communication*. London: Longman.

Gardner, R. C., and Lambert, W. E. (1972). *Attitudes and Motivation in Second Language Learning*. Rowley, MA: Newbury House.

Gathercole, S. E. and Baddeley, A. D. (1993). *Working Memory and Language*. Hillsdale, NJ: Erlbaum.

Globerson, T. (1986). When Do Structural Changes Underlie Behavioral Changes? In I. Levin (Ed.), *Stage and Structure: Re-opening the Debate*. Norwood, N. J.: Ablex Press.

Harrington, M. (1992). Working Memory Capacity as a Constraint on L2 Development. In R. J. Harris (Ed.), *Cognitive Processing in Bilinguals*. Elsevier, Netherlands: North Holland.

Hakuta, K. (1976). Becoming Bilingual: A Case Study of a Japanese Child Learning English. *Language Learning, 26*, 321-351.

Hasher, L. and Zacks, R. T. (1979). Automatic and Effortful Processes in Memory. *Journal of Experimental Psychology: General, 108*, 356-388.

Hatch, E. (1978). Discourse Analysis and Second Language Acquisition. In E. Hatch (Ed.), *Second Language Acquisition: A Book of Readings*. Rowley, MA: Newbury House.

Hermann, G. (1980). Attitudes and Success in Children's Learning of English as a Second Language: The Motivational Versus the Resultative Hypothesis. *English Language Teaching Journal, 34*, 247-254.

Heyde, A. (1977). The Relationship Between Self-Esteem and the Oral Production of a Second Language. In H. D. Brown, C. Yorio, and R. Crymes (Eds.), *On TESOL '77. Teaching and Learning English as a Second Language: Trends in Research and Practice*. Washington, TESOL.

Howard, D. V. (1983). *Cognitive Psychology: Memory, Language and Thought*. New York: Macmillan.

Huebner, T. (1983). *A Longitudinal Analysis of the Acquisition of English*. Ann Arbor: Karoma Publishers.

Karmiloff-Smith, A. (1986). Stage/Structure Versus Phase/Process in Modelling Linguistic and Cognitive Development. In I. Levin (Ed.), *Stage and Structure: Re-opening the Debate*. Norwood, N. J.: Ablex Press.

Kolers, P. A., and Roediger, H. L. (1984). Procedures of Mind. *Journal of Verbal Learning and Verbal Behavior, 23*, 425-449

LaBerge, D. and Samuels, S. J. (1974). Towards a Theory of Automatic Information Processing in Reading. *Cognitive Psychology, 6*, 293-323.

Levelt, W.J.M. (1977). Skill Theory and Language Teaching. *Studies in Second Language Acquisition, 1*, 53-70.

Lewis, C. (1981). Skill in Algebra. In J. R. Anderson (Ed.), *Cognitive Skills and Their Acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lightbown, P. M. (1985). Great Expectations: Second-Language Acquisition Research and Classroom Teaching. *Applied Linguistics, 6*, 173-189.

Long, M., and Sato, C. J. (1984). Methodological Issues in Interlanguage Studies: An Interactionist Perspective. In A. Davies, C. Criper, and A. P. R. Howatt (Eds.), *Interlanguage*. Edinburgh: Edinburgh University Press.

McGhie, A. (1969). *Pathology of Attention*. Baltimore: Penguin.

McGroarty, M. (1989). *The "Good Learner" of English in Two Settings*. University of California, Los Angeles: Center for Language Education and Research.

McLaughlin, B. (1980). Theory and Research in Second-Language Learning: An Emerging Paradigm. *Language Learning, 30*, 331-350.

McLaughlin, B. (1990). Restructuring. *Applied Linguistics, 11*, 1-16.

McLeod, B. and McLaughlin, B. (1986). Restructuring or Automaticity? Reading in a Second Language. *Language Learning, 36*, 109-123.

Naiman, N., Frohlich, M., Stern, H. H. and Todesco, A. (1978). *The Good Language Learner*. Toronto: The Ontario Institute for Studies in Education, Rowley, MA: Newbury House, 1972.

Nation, R. and McLaughlin, B. (1986). Experts and Novices: An Information-Processing Approach to the "Good Language Learner" Problem. *Applied Psycholinguistics, 7*, 41-56.

Nayak, N. Hansen, N., Krueger, N. and McLaughlin, B. (1990). Language-Learning Strategies in Monolingual and Multilingual Subjects. *Language Learning, 40*, 221-244.

Nelson, R., and Jakobovits, L. A. (1970). Motivation in Foreign Language Learning. In J. Tursi (Ed.), *Foreign Languages and the New Student: Reports of the Working Committees*. New York: Northeast Conference on the Teaching of Foreign Languages.

Newell, A. (1973). Production Systems: Models of Control Structures. In W. G. Chase (Ed.), *Visual Information Processing*. New York: Academic Press.

Oller, J. W. (1981). Research on the Measurement of Affective Variables: Some Remaining Questions. In R. W. Andersen (ed.), *New Dimensions in Second Language Acquisition Research*. Rowley, MA: Newbury House.

O'Malley, J. M., and Chamot, A. U. (1989). *Learning Strategies in Second Language Acquisition*. New York: Cambridge University Press.

Oxford, R. L. (1986). *Second Language Learning Strategies: Current Research and Implications for Practice*. University of California, Los Angeles: Center for Language Education and Research.

Pike K. (1960). Nucleation. *Modern Language Journal, 44*, 291-295.

Posner, M. I., and Snyder, C. R. R. (1975). Attention and Cognitive Control. In R. L. Solso (Ed.), *Information Processing and Cognition: The Loyola Symposium*. Hillsdale, NJ: Erlbaum.

Ramsey, R. M. G. (1980). Language-Learning Approach Styles of Adult Multilinguals and Successful Language Learners. *Annals of the New York Academy of Sciences, 345*, 73-96.

Rumelhart, D. E., and Norman, D. A., (1978). Accretion, Tuning, and Restructuring: Three Modes of Learning. In Cotton J., Klatzky R. (Eds.). *Semantic Factors in Cognition*. Hillsdale, N.J: Lawrence Erlbaum Associates.

Schneider, W. and Shiffrin, R.M. (1977). Controlled and Automatic Human Information Processing: I. Detection, search, and Attention. *Psychological Review, 84*, 1-66.

Selinker, L. (1972). Interlanguage. *IRAL, 10*, 209-231.

Service, E. (1992). Phonology, Working Memory, and Foreign Language Learning. Quarterly Journal of Experimental Psychology: *Human Experimental Psychology, 45*, 21-50.

Shiffrin, R. M., and Schneider, W. (1977). Controlled and Automatic Human Information Processing: II. Perceptual Learning, Automatic Attending, and a General Theory. *Psychological Review*, 84, 127-190.

Speidel, G. E. (1993). Phonological Short-Term Memory and Individual Differences in Learning to Speak: A Bilingual Case Study. *First Language*, 13, 69-91.

Strong, M. (1984). Integrative Motivation: Cause or Result of Successful Second Language Acquisition? *Language Learning*, 34, 1-13.

Wenden, A. L. (1987). Metacognition: An Expanded View on the Cognitive Abilities of L2 Learners. *Language Learning*, 37, 573-598.

Wong Fillmore, L. (1976). *The Second Time Around: Cognitive and Social Strategies in Second Language Acquisition*. Doctoral Dissertation, Stanford University.

IMPROVING THE MEASUREMENT OF LANGUAGE APTITUDE: THE POTENTIAL CONTRIBUTION OF L1 MEASURES

John W. Thain
Defense Language Institute

Overview

Background

This paper begins with a brief sketch of work done in the area of language aptitude measurement at the Defense Language Foreign Language Center (DLIFLC) in the past eight years. There is no effort to go into detail into this sketch; however, the reader interested in further detail is provided with ample references to other presentations at this symposium and to other published works in the footnotes and bibliographic references following this paper. This cursory introduction does, however, define the instruments and measures used to screen potential students applying for language training at DLI. The intent is that this sketch will help provide context and points of orientation for the reader later on in this paper.

Adding L1 Measures as Predictors

The rest of this paper addresses the feasibility of adding two specific L1 measures as additional predictors to the current Defense Language Aptitude Battery (DLAB). DLAB is one of the batteries used to screen applicants for language training at DLI. The two potential predictors are (1) a test of L1 (native-language) listening comprehension and (2) a test of sensitivity to English grammar and usage.

Most of the paper deals with only one of the two potential predictors, an L1 measure of listening comprehension. The second potential predictor, a test of sensitivity to English grammar, is discussed only briefly.

The Main Body of the Paper: L1 Listening Comprehension as a Predictor

Five sections on listening comprehension in this paper

The part of the paper dealing with native-language listening comprehension can be further subdivided into five sections. The first section reviews the kinds of native listening comprehension (NL) tests currently available as models. The next three sections address several theoretical issues involved in the addition of NL tests to the current DLAB. The last section lists conclusions and recommendations.

Importance of the middle three sections

The content of the middle three sections mentioned above deserves further comment. There is little precedent for using NL tests as foreign language aptitude tests. A literature search was needed to address the relevant theoretical issues in using NL tests. I found three approaches in the literature that were relevant to the question of using NL tests as language aptitude predictors. Each approach was represented by its own literature, but no previous attempt had been made to synthesize information from these three perspectives to address the specific problem at hand. I call the three perspectives (1) the predictive perspective (2) the linguistic content perspective, and (3) the perspective of cognitive models. I needed not only to review three different kinds of literature, but in a sense, to attempt an unprecedented synthesis of three types of literature for a particular purpose. Hence, there needed to be three middle sections under the general topic of NL comprehension, each section concerned with one of the three approaches.

The First of Three Approaches in the Literature on Listening: the Predictive Perspective

I call the first approach the predictive perspective. In the section dealing with this perspective, I refer to studies of the statistical characteristics of currently used screening measures, and the potential consequences for overall prediction of adding additional predictors. I address the effect of covariance between predictors on the total predictive power of a battery. I also mention the consequences of adding predictors that may themselves be multidimensional to existing predictors in a battery.

The Second of Three Approaches in the Literature on Listening: the Linguistic Content Perspective

I call the second approach the linguistic content perspective. The discussion of this perspective is more lengthy and complex than the discussion of the other two perspectives.

I point out that FL (foreign language) listening proficiency is one of the proficiency criteria we want to predict. I note how the concept of language proficiency in all skills, including listening, as expressed by the Interagency Language Roundtable (ILR) proficiency level scale, has been influenced by very basic, important, and overwhelmingly positive theoretical developments in the field of foreign language teaching methodology over the years. In the course of these developments, the ILR proficiency levels have established their unquestioned legitimacy as training criteria within the government and a large part of the progressive academic teaching community.[1]

Two consequences of the broad range of ability encompassed in ILR scales.

The ILR listening scale attempts to quantify a very broad range of proficiency. The lower part of the scale describes beginning language learners and the upper part of the scale describes polished bilinguals. This enormous range of individual differences seems to bring about two consequences.

Consequence number one. The first consequence is that different aptitude predictors may represent abilities that contribute in different magnitude at different levels of proficiency acquisition (and thus at different points on the ILR listening scale). I also note that the listening literature suggests there may be two types of listening, and that these two types of listening may make different cognitive demands on the listener. Each type of listening may have its own unique pattern of relationships with the other ILR skills. I conclude that evidence of multidimensionality in listening and of complex interrelationships among ILR skills could have interesting consequences for predictor-criterion relationships.

Consequence number two. A discussion of the first consequence leads us naturally to the second consequence. The ILR scale is "a "vertical" scale rising from Levels 0 to 5, a very great range of the ability. NL research looks at listening from a "horizontal" view that intersects only the top of the vertical ILR scale. Factors such as grammar, vocabulary, and phonology that play a major role for beginning FL listeners play a much lesser role in NL. In turn, NL research has identified separate listening factors that contribute to individual differences among native listeners, and these factors do not correspond to the factors contributing to individual differences at lower levels on the ILR scale.

Pure traits (PTs) vs. native authentic listening (NAL)

The difference between the "vertical" and "horizontal" perspectives is highlighted as I cite the work of the NL researchers Bostrom and Waldhart (1981). They resolved NL into three factors: (1) short-term listening (2) long-term listening (3) interpretive listening (sensitivity to affect).

---

[1] There are other scales for rating proficiency that are based on level systems similar to that used by the ILR. Examples include the ACTFL scale used by the American Council of Teachers of Foreign Languages and other rating scales used in Europe.

I contrast a view of NL based on a three-factor analysis similar to that of Bostrom and Waldhart, and a "global" view of NL as "native authentic listening" (NAL). After I coin a term by calling each factor in the three-factor analysis a "pure trait" (PT), I broach an important question that I do not immediately answer: "Is a NL test based on PTs a better predictor of ILR proficiency levels than a test based on NAL?"

The perspective of cognitive models

I call the third perspective the cognitive modeling approach. I sketch evolutionary changes in the field of psychology from radical "black-box behaviorism" days to current day cognitive psychology, including the development of the field of artificial intelligence (AI). AI specialists have successfully modeled human comprehension of language. Within a limited range of topics, machines can now carry on reasonable conversations with humans in which they make many of the inferences that humans would make in similar circumstances.

In the context of these developments in AI, I draw a series of analogies between listening comprehension and the operation of a multimedia database. I point out that the series of analogies leads to conclusions similar to those of Bostrom and Waldhart concerning the multidimensional nature of NL.

Conclusions and recommendations for further study about listening comprehension

The last section on listening comprehension lists conclusions and recommendations. I list a set of criteria for evaluating possible listening comprehension measures for inclusion into the DLAB. I categorize the NL tests reviewed earlier in terms of whether they measure PTs (pure traits), native authentic listening (NAL), or some mixture of the two. I then list some of the issues in using PTs as language aptitude measures, and related issues in using measures of NAL as language aptitude measures.

The Rest of the Paper: Tests of Grammatical Sensitivity as Predictors

In the last part of this paper, I review tests of grammatical sensitivity, but not in the same detail with which I reviewed NL tests earlier. Two types of tests are reviewed: (1) tests of sensitivity to English grammar, and (2) tests of sensitivity to foreign (or artificial) language grammar rules.

Overview of organization of the paper

- This overview spans pages 1-3 of the paper.
- A sketch of background information and references to related presentations at this symposium are to be found at pages 4-6.
- A major division of the paper entitled "Exploring Native Listening Comprehension" spans pages 6-28.
    - A review of currently available NL comprehension tests is found at pages 6-7 under the main division heading.
    - A section entitled "First of Three Complementary Approaches: the Predictive Perspective " spans pages 8-9.
    - A section entitled "Second of Three Complementary Approaches: the Linguistic Content Perspective "spans pages 9-16.
    - A section entitled "Last of Three Complementary Approaches: the Predictive Perspective:" spans pages 17-23.
    - A section entitled "Conclusions and Recommendations Concerning NL Measures" spans pages 23-28.
- A major division of the paper entitled "Exploring Tests of Grammatical Sensitivity in English" covers pages 29-32.
- Bibliographic references are found at pages 32-38.

Background Information and Related Presentations at this Symposium

General

A major study conducted at the Defense Language Institute (DLI) from 1986 to 1989 investigated how well a variety of variables predicted proficiency after language training. This study, the Language Skill Change Project (LSCP), was a longitudinal study designed to follow approximately 2,000 Army "linguists" throughout a four-year period. Data collection points included (1) initial aptitude screening prior to entry into the Army; (2) several occasions in the course of language training,; and (3) post-graduation field assignments. The population sample included both students of Spanish, German, Russian, and Korean. Another presentation describes this population sample in more detail.

A secondary study used a portion of the same data base to investigate predictors of attrition from DLI training.

In both the longitudinal and the attrition studies, the predictor variables used to predict language training success included (1) scores on a general vocational aptitude battery and a language aptitude battery, both used to screen potential students; (2) scores on other cognitive measures not used in the screening process; and (3) scores and ratings on measures of student motivation, anxiety, and use of learning strategies.

Criterion measures included (1) successful course completion as opposed to attrition from training; and (2) the Defense Language Proficiency Tests in these languages for speaking, listening, and reading skills.

Aptitude Variables used in Official Screening

Aptitude tests used in official screening are not administered by the DLI. These tests are normally administered by the interservice Military Enlistment and Processing Command (MEPCOM).

Applicants for military service must attain passing scores on a composite of the Armed Services Vocational Aptitude Battery (ASVAB), a paper and pencil general vocational aptitude battery. The passing scores have hardly changed since the LSCP was conducted. ASVAB includes tests of verbal, mathematical, technical (mechanical and electrical), and clerical coding abilities. The verbal tests include measures of paragraph comprehension and vocabulary knowledge. All ASVAB test materials are printed in English.

Examinees reaching certain minimum scores in specified components of the ASVAB are eligible to take the Defense Language Aptitude Battery (DLAB). This battery contains several subtests.[2] The subtests measure (1) identification of syllable stress (2) deductive language learning of an artificial language (3) inductive language learning from pictures and artificial language work sample. The first two subtests are presented on tape, and the third subtest is printed in the test booklet.

Scores on the ASVAB and DLAB tests administered by MEPCOM would normally be present in the official personnel records of students even before students arrive at DLI.

Other Cognitive Measures Not Used in Official Screening

After completing basic training, the students in the LSCP sample actually arrived at DLI. DLI administered additional cognitive tests to them as part of the LSCP. These tests included the Watson-Glaser Critical Thinking Appraisal, the Flanagan Expression Test, and the Flanagan Memory Test.

---

[2]See reference by Petersen, C., Al-Haik, A. (1976)

Measures of Student Motivation, Anxiety, and Learning Strategies

In order to assess motivation to learn a foreign language immediately prior to language training, the subjects were administered Gardner Questionnaire Form A. This questionnaire was a modification of previous questionnaires used by Gardner in earlier research and included scales for Integrativeness, Instrumental Motivation, and Interest in Foreign Language. [3]

During the course of language training, Gardner Questionnaire Form B was administered. This questionnaire included scales for Motivational Intensity, Attitude Toward Learning, Class Anxiety, Use Anxiety, Desire to Learn, Attitude Toward the Instructor, Attitude Toward the Course.

The Strategy Inventory for Language Learning (SILL) was also administered to measure self-reported use of learning strategies during instruction.

Results of Background Studies

The results of these studies have been already described in another paper at this symposium.[4]

In the basic study, stepwise multiple regressions with forced order of entry indicated that (1) general vocational aptitude (measured by ASVAB), (2) language-learning aptitude as (measured by DLAB), (3) measures of student motivation, anxiety, and learning strategies use, (4) additional cognitive measures not included in the official screening process all added contributions to predictive power. However, the pattern of multivariate prediction varied across the four languages taught and across the three criterion language skills .

A secondary study used a restricted set of variables. Course completion (as opposed to attrition from training) was used as a criterion measure. Chi-square interaction analyses (CHAID) indicated that (1) the pattern of interaction of variables varied across languages (2) both DLAB and the additional cognitive measures not included in the screening process contributed to the segmentation of subsamples. The subsamples in individual languages were segmented on the basis of the differentiating criterion of percentage of successful course completion.

Related studies and follow-up studies

Shortly after the above mentioned studies were completed, DLI launched several simultaneous efforts to improve aptitude prediction: (1) an item analysis of the current DLAB (2) an effort to compare languages in terms of the "factors" that made some languages more difficult to learn than others (3) an effort to specify the kinds of language abilities and measures that should be included in an aptitude battery.

The results of the item analysis of DLAB were reported in another presentation at this symposium.[5]

Another presentation at this symposium addressed the second and third efforts mentioned above.[6]

---

[3] See reference by Gardner, R., Lalonde, R., Moorcraft, R., Evers, F. (1985).

[4] "The Defense Language Aptitude Battery: What is it and how well does it work?", by John Lett and John Thain.

[5] "The Defense Language Aptitude Battery: What is it and how well does it work?", by John Lett and John Thain.

[6] "Psycholinguistic Issues in the Assessment of the Subcomponents of Language Abilities, by Brian MacWhinney.

Conclusions drawn concerning possible addition of L1 measures to DLAB

As noted above, the current DLAB contains test items based on artificial language material. This material taps primarily grammar learning and grammar analysis abilities. It does not contain test material based on normal L1 (English) language.

The other battery used in official screening process; the ASVAB, does include written L1 (English) tests of verbal ability, but does not include auditory tests.

DLI staff examined all of the information from the LSCP data base and recommendations resulting from the follow-on work mentioned above. DLI then decided to explore the possibility of adding two additional predictors to the language aptitude battery: (1) an L1 native speaker (English) test of listening comprehension (2) a test of sensitivity to English grammar and usage.

Exploring Native Language Listening Comprehension

Review of native-language (NL) listening tests

Introductory comments

I began my exploration of L1 native listening comprehension as a potential predictor by reviewing English native listening (NL) tests.

I discovered that NL test developers tended to see NLs as listening "skill-users" with a function and corresponding work to do in the native society. These developers perceived the NL as a student, teacher, counselor, or businessman; they felt his function was to learn, to help others, or to serve as an employer. NL test developers differ from FL test developers in this respect. They show less interest in clearly separating "language listening skills" from other useful skills and knowledge.

I quickly detected something interesting about English listening comprehension testing of *foreign students at English-speaking universities*--namely the tests used had more in common with NL tests of listening than with tests of foreign language (FL) listening comprehension. For this reason, we included such listening tests in our review.

I also found another interesting difference between contemporary NL and FL listening testing and research. Nowadays many FL testers, especially those at federal government institutions, want to test "proficiency," i.e. authentic and useful language. They don't want to test anything that looks like a classroom drill or an isolated piece of language. On the other hand, NL researchers are showing interest in testing memory span for letters and similar tests of short term memory.

While NL testers may concede such skills may be not useful in isolation, they tend to find these measures to be useful as (1) predictors of more complex behavior, or (2) moderating variables for cognitive models of more complex skills, or (3) diagnostic devices.

NL Tests Reviewed

I reviewed seven tests which I found to be mentioned in the literature. Brief synopses follow:

Watson-Barker Listening Comprehension Test. This test includes subtests for "listening to a lecture," "emotional listening," "instructions and directions," "listening for content," and "listening to conversations." Businesses have used this test to accompany training programs. The University of Illinois has used it to differentiate levels of listening skills of foreign students taking classes at the University. The test is presented by means of videotape. The publisher is Spectra Communications in New Orleans, LA.

Kentucky Comprehensive Listening Test. This test is a multiple choice test with four parts. The four parts measure performance on the following tasks:

(1) Listening to letters or number strings amidst distracting noise. The examinee is prompted immediately after the stimulus to identify the relative position of a letter or number in the string.

(2) Listening to letters and number strings without the presence of distracting noise, but with a delayed prompt to identify the relative position of a letter or number.

(3) Listening for real meanings (i.e. illocutionary acts) hidden in very short answers in a dialogue with strong nonverbal affective signals

(4) Listening to a 1500 word lecture.

The publisher is the Kentucky Listening Research Center in Lexington, Kentucky. Data collected on this test are particularly interesting (1) because it has been used in a variety of research and practical contexts, (2) the authors have fostered a series of studies from which a particularly fruitful nexus of explanatory constructs has evolved.

Carleton University Test.[7] Carleton University in Ottawa, Ontario, has constructed a listening test that it administers to its incoming foreign students. The examinees take notes on a lecture, actually reorganize their notes, and then do library research on the basis of their reorganized notes. The criterion for success is the quality of their library research. Test results reportedly correlate highly with an English comprehension test developed by the University of Michigan.

NTE Core Battery Test of Communication Skills. The National Teacher's Examination (NTE) program includes a Test of Communication Skills, which includes subtests in listening, reading, and writing. Many sample listening items given in the test information brochure are based on typical listening comprehension situations. However, item content is biased toward typical situations in which teachers might be involved. Some of questions defining the examinee's task include: (1)"Why does the man hesitate to call William's parents?;" and (2) "What assumption does the speaker make about high schools?"

The NTE School Guidance and Counseling Examination. This test includes a listening component, which is administered as part of a larger battery. The battery as a whole evaluates the skills and knowledge required of school counselors. In this test, thc examinee listens to test items depicting situations in which counselors may be involved. The examinee then answers multiple-choice items introduced by item stems such as "The client is likely to react by..." or "The counselor's objective was..."

Brown Carlsen and STEP. Two older NL tests include (1) the Brown-Carlsen Listening Comprehension Test, from Harcourt Brace and Jovanovich, and (2) the STEP (Sequential Tests of Educational Progress) Listening Comprehension Test, once published by a since dissolved ETS subsidiary. The Brown-Carlsen test has subscales that measure vocabulary, recognition of transitions, ability to follow directions, immediate recall, and retention of facts from a lecture. The STEP listening test was one of seven tests in a battery, which included tests of reading, spelling, and other achievement areas. It was published in a series of forms that spanned grade levels 4-14.

Introducing Three Complementary Approaches in the Literature on Listening Comprehension

In a general sense, there is an abundance of literature on NL. On the specific point of view of use NL as a predictor of foreign language proficiency, there is a poverty of literature.

The general literature on NL suggested several complementary perspectives for understanding the subject area. I became aware that many people in the field of language aptitude measurement may seldom have considered these perspectives about NL *in conjunction with each other*. I believe the approaches are

---

[7] Personal communication from Janna Fox at Carleton University. See also reference by Janssen, C., Hansen, C., Buck, G., DesBrisay M., Fox, J., Shohamy, E., (1993).

synergetic. This means that insights and conclusions gained from one perspective can influence one's thinking in following up other approaches. One of my objectives is to improve communication between investigators using different approaches and to stimulate discussion about new ideas arising from the interaction of approaches.[8] I will first touch on several seemingly loosely related ideas, and then attempt to tie them together with some concrete examples.

I have called three of these diverse points of view the (1) predictive perspective; (2) the linguistic content perspective; and (3) the cognitive model perspective.

First of Three Complementary Approaches: the Predictive Perspective:

The general standard regression formula for prediction is:

$$Y = \sum_{i=1}^{n} \alpha_i X_i + C; \quad \text{where } \alpha_i X_i \neq 0, \quad n \geq 1.$$

In this general formula, Y is the criterion, and i is the number of predictors contributing to the equation. Each of the n predictor values is multiplied by its own weight α_i and then all the weighted predictors are summed to give the overall weighted contribution of all the predictors in the equation. The values of the weights are affected by covariance between the predictors. This general formula can apply to the prediction of any proficiency criterion from any number of NL predictors.

The mathematics of prediction are straightforward. However, communication problems can arise among investigators with different backgrounds for reasons that have little to do with the mathematics of prediction. For this reason, in the following paragraphs I will be trying to accomplish two things at once. I will list the possible predictors that might go into a predictive equation, but at the same time I will also be explaining how researchers with different perspectives might have divergent views on how many predictors should be in the equation, and how these predictors are interrelated.

L1 Predictors already included in general aptitude batteries even before language aptitude testing.

In the case of the Defense Language Institute, a passing score on a general aptitude battery, the ASVAB, is a prerequisite for taking the DLAB. Hence, there are already some potential predictors from ASVAB available for inclusion in the equation above (before considering any specific FL aptitude predictors or any new potential L1 predictors.) General aptitude tests such as the ASVAB typically include subtests that represent the V (Verbal) factor as well as other familiar factors such as the N (numerical) factor.[9]

---

[8] Some of these approaches may seem on the surface to diverge from the ideas underlying our use of the ILR proficiency scale as criteria. Where this may seem to be the case, I will pause to explain exactly what elements of these approaches I find useful and compatible with the ILR approach.

[9]Other factors in ASVAB (or similar general aptitude measures) besides the V factor are likely to contribute to the prediction of language proficiency. The V factor is mainly relevant to the discussion here in this section, because this section focuses on L1 comprehension measures. For more detail, see references by Silva, J., White. L. (1992); Department of Defense (1985); Kass, R., Mitchell K., Grafton, F., Wing, H. (1983); Carroll, (1958); Carroll (1962), Carrol (1993). Tests that consistently correlate with each other more than with other types of tests are assigned to the same "factor". The "V" factor is consistently represented by L1 vocabulary tests. There is no hard and fast theoretical reason in the field of psychology that a "V" factor should be exclusively identified with any of the four L1 skills. However achievement and aptitude batteries, including ASVAB, normally include a reading comprehension test; but they include tests of the other three skills less often. It is easier to produce, administer, and score multiple-choice reading

Additional predictors in language aptitude batteries not identifiable with any of the four skills

A variety of studies have identified factors related to phonology, grammatical sensitivity, and word association that contribute independent variance toward the prediction of L2 Proficiency beyond that contributed by the "V" L1 factor included in general aptitude batteries. However,-- (1) the "V" factors, (2) these additional aptitude factors, and (3) any additional L1 predictors we may choose to add--may *all* share some covariance. This covariance would (1) affect the weights in the prediction equation so that all weights would have to be recomputed with the addition of each new predictor, and (2) tend to limit increases in the size of a multiple correlation coefficient with the addition of each predictor, (to the extent that each predictor added shared variance with predictors introduced earlier in the equation.)

Research on Potential L1 Listening Predictors that lack parallelism to ILR/ACTFL criterion scales

FL researchers using the ILR and ACTFL proficiency scales as criteria tend to consider L2 listening as a unitary trait. They may tend to assume that NL listening would also be an unitary trait. If NL listening were a unitary trait, a single additional predictor would be added to the equation to join the predictors mentioned earlier.

However, a contrasting perspective will be discussed later in this paper. At that time, I will point out that two prominent NL researchers have attempted to analyze NL into three component traits. Users of the ILR/ACTFL scale may be forewarned that only one of these traits bears some similarity to the kind of global "listening" with which they are familiar. My interest is in these NL component traits as potential NL *predictors* of ILR proficiency levels, *not* as *alternatives* to the ILR listening proficiency *criteria*.

Concepts of FL listening that are different in emphasis from the ILR/ACTFL perspective

In addition to the NL researchers discussed in the previous paragraph, there are some writers on FL listening who at times don't see the four language skills as distinct "points," as much as moist blurry ink blots that overlap each other. My interest is in how their insights shed light on the aptitude-proficiency (*predictor-criterion*) correlations across languages and language skills. The purpose is not to advance these ideas as alternatives to the ILR skill level *criteria.*

In the next section on the "linguistic content perspective," I will attempt to explore predictor-criterion relationships from a different point of view and attempt to bridge a communication gap between researchers with different points of view.

Second of Three Complementary Approaches in the Literature on Listening Comprehension: the Linguistic Content Perspective

Introduction.

There is another reason why foreign language listening and NL researchers might not initially communicate. NL researchers may not be very familiar with the development of FL teaching methodology in the past 75 years. For this reason, I will very briefly sketch how the relevant developments in teaching methodology may have shaped the ILR listening scale and ILR testing procedures. The intent is to establish the critical importance of the ILR scale as a foreign language criterion measure.

I then discuss alternative conceptualizations of the interrelationships among FL listening and other FL skills advanced by scholars not closely associated with the ILR testing community. The intent is to add

---

tests than tests of the other three skills. Hence psychologists tend to identify L1 reading comprehension almost as closely with the "V" factor as vocabulary tests.

a relevant perspective to our in-house thinking about skill prediction. Finally I contrast a generally accepted concept of FL listening with a "nonparallel" concept of "listening" advanced by two NL researchers.

From Grammar Translation to Proficiency

In 1930, the grammar-translation method was used to teach foreign languages in America. By 1960, the audiolingual method had displaced this earlier method. Since then, specialists in FL teaching methodology have been distancing themselves from both approaches.

They have begun to define authentic language use is the *ultimate* instructional goal. That is, the ultimate goal (perhaps unattainable by most second language learners) was that second language learners should read, speak, listen, and write languages the way native speakers use their language. In addition, the FL methodologists recognized and defined a variety of *intermediate* levels of ability for using and understand authentic language short of native speaker capability.

The FL methodologists rejected the grammar translation approach because they noticed that second language learners could learn grammar rules and do translations without much progress toward being able to read, speak, listen, and write languages the way native speakers used the language (or toward any recognized intermediate level of authentic language use).

They also rejected the audiolingual approach because they noticed that although second language learners could acquire habits for listening to and repeating small segments of language, they were not necessarily making progress in the sense of using progressively more complex cognitive processes through the new language.

The Proficiency Movement

Prominent FL methodologists joined together in the "proficiency movement." This movement included representatives from the structured, intensive language programs of the Federal Government and from a variety of language programs within academic institutions. Members of this movement began to define proficiency as their criterion goal. They defined a proficiency scale for each skill in terms of increasing ability to accurately use the new language to accomplish increasingly difficult authentic language tasks. All foreign language learners as well as native speakers were rated on a continuous scale across an enormous range of ability, from rank beginners to polished bilinguals--students and teachers alike.[10] Testing tasks and items showed a corresponding range of difficulty. At any given point on this broad continuum of item difficulty, it was assumed that an item on a listening proficiency test should be set in a meaningful situation in which language students might actually find themselves using the language in real life.

Today a progressive ILR tester in the proficiency movement may tend to consider it a throwback to obsolete unproductive methodology to include items in a FL listening test which consist of isolated bits of language (for example, isolated sounds or letters). Furthermore, such a tester might argue that a test of

---

[10] It is very important that there be a cooperative program between the government and academia to use compatible testing systems that measure such a broad range of ability. A teacher needs to master the language he teaches, and also master the language in which his/her employing institution imparts training in FL methodology (usually English in the United States). For these reasons, it is difficult to conceive an effective national level policy for fostering language training in this country without such a testing system. A testing system is needed to manage the career cycle of the two main classes of people who become foreign language teachers in America: (1) American-born language students who learn enough of a foreign language to be able to themselves teach foreign languages to other Americans; and (2) foreign-born teachers who first become students to learn English and then subsequently teach foreign languages to Americans. For more detail see references by Carroll, J. (1967); Higgs, T., Clifford R. (1982); Heileman, L., Kaplan, I., (1985), James, C. (ed.), (1985), Lowe, P. (1985), Clark, J. (1986), Child, J. (1987), Valdman, (ed.), (1987), Clark, J., Clifford R. (1987), Child, J., Clifford, R., Lowe, P. (1993), Hadley, A. (1993).

listening should not permit the examinee to answer a question or solve a communication problem without being forced to understand the lexis and grammar of a foreign language text (e.g. as would be the case if the examinee answered a question about text solely by correctly interpreting a combination of gestures and voice modulation or by relying on background knowledge and context.)

Effect of Unrestricted Range in the Population Tested on Observed Variance

When we FL researchers define a "listening" trait using a rating scale like the ILR scale for a broad population ranging from beginning learners to polished bilinguals, we find statistical evidence for considering "listening" a unitary trait. An overwhelming amount of variance is contributed by huge individual differences in mastery of foreign language codes. All other possible contributing traits are but drops in this vast ocean of variance.

The analogy of a vast ocean and vast variance can be extended further. ILR proficiency scales depend on individual differences in factors such as vocabulary, grammar, and sociolinguistic competence to discriminate among a great range of ability in the population.[11] The situation may be different for NL testing. In contrast, many differences in native listener performance may be less dependent on individual differences in vocabulary, grammar of the native language or knowledge of one's own native culture than on other factors. This suggests a way to complete the ocean analogy. If somehow all the water in the ocean evaporated, a theory based on the ocean being comprised of 96% water would not be a very good schema for making an inventory of the salts, minerals, fish, plants, and rocks left behind.

Good for the goose, but not for the gander

I hesitate to consider my experiences as an ILR proficiency tester as a warrant to evaluate the kind of issues that should be considered important in the field of NL testing (or specifically in the field of NL testing as a predictor for FL proficiency.) In this area, I believe ILR testing experience needs to be supplemented by perspectives from NL testing, and by other perspectives from the FL research community.

However, before proceeding to introduce some other helpful and complementary perspectives, let me hasten to preclude any misunderstanding based upon my previous statements. In general and for all practical purposes, I consider (1) that FL teaching methods have evolved in the right direction; (2) the concomitant trend toward accountability both in the government and in universities is good; and (3) our ILR criterion of "foreign language proficiency," specifically including listening proficiency, is defined properly.

An overview of other perspectives from the FL research community

Should skills be viewed as "distinct points" or "blurry inkblots?" Table 1 lists distinctions found in the literature that potentially cut across skills. The information in the table highlights the possibility that some types of L1 listening may make cognitive demands that are similar to those required in L1 speaking, while other types of L1 listening may make cognitive demands that are more like L1 reading.

If we plan to use L1 listening to predict L2, these distinctions are potentially important because the distinctions in L1 may have parallels in L2. Tannen's (1982) oral-literate style distinction may illustrate this point

---

[11] For a more complete elaboration of this point, see reference by de Jong, J. (1994). As de Jong points out, if one looks *closely* at *any* narrow subinterval on the broad scale of language proficiency (not just at the top of the scale for native proficiency as I am doing in this paragraph), one can probably find evidence for trait multidimensionality within that specific *subinterval.* On the other hand, if one takes a *broad* overview of the *whole* language proficiency scale (from a *distance* to use de Jong's metaphor), the scale as a whole appears to be unidimensional.

TABLE 1[12]
TWO TYPES OF LISTENING?
A CLASSIC EXAMPLE OF FUZZY SETS

SOURCE	MORE LIKE SPEAKING	MORE LIKE READING
Tannen (1982)	Oral style	Literate style
ILR	Street	School
ILR	Participatory	Nonparticipatory
Bostrom (1981)	Interpretive listening	Lecture listening
Cummins (1982)	Contextualized Requires BICS (Basic Interpersonal Communication Skills)	Decontextualized Requires CALP (Cognitive Academic Language Proficiency)
Canale (1982)	Interactive	Autonomous
Rost (1990)	Collaborative	Transactional
MBTI Thinking/Feeling[13]	Feeling type favored	Thinking type favored
Brain-hemisphere studies	Right brain favored	Left brain favored
Other	Situation-based	Idea-based
	Listener plans to politely clarify speaker's role, intentions, or feelings as part of listening process.	Listener plans to make mental or written notes as part of listening process, with the intention of later consulting dictionaries, textbooks, or other reference works.

Some measures of L1 listening may (1) be more closely related to L1 reading; (2) tend to covary with ASVAB, because ASVAB as a whole is probably more "literate" than "oral;" (3) tend to predict L2 listening skills that are more "literate" than "oral".

Other measures of L1 listening may (1) be more closely related to L1 speaking (2) tend to add distinct variance not already represented in ASVAB (3) tend to predict L2 listening skills that are more "oral" than "literate."

The above observations seem to have potential predictive consequences: (1) adding L1 listening predictors may improve prediction of other ILR skills than listening as much or more than these predictors

---

[12] It should be emphasized that the two types of listening implied by Table 1 above are classic examples of "fuzzy sets." The various distinctions listed cut across each other and overlap. For example, (1) some lecturers may use "oral" styles to better communicate technical information to their audience (2) some face-to-face speakers may address very technical or even esoteric subjects. (3) certain lecture and staff meeting settings may be viewed as continuous discourses in which the listener shifts back and forth from a nonparticipatory status to a participatory status (as in question and answer sessions after lectures, or in briefings from individual departments in the course of some staff meetings) (4) certain interactive situations could place demands on the "thinking," "left brain", "idea-oriented" side of the listener, while certain noninteractive situations could place demands on the "feeling", "right brain", and "people-side" of the listener. Although the list of fuzzy points admittedly could be extended indefinitely, I still think there seems to be enough of a pattern present to talk about two "fuzzy sets" rather than a list of totally random and unrelated distinctions.

[13] See reference by Myers, J., McCauley, M. (1985).

may improve prediction of listening itself (2) it may not be possible (or even desirable) to have a neat paradigm of L1 skill predictors that match corresponding L2 skill criteria.

A perspective for viewing predictor-criteria interactions. Figure 1 portrays the kind of predictor-criterion relationships suggested in the previous section. The right and center portions of the diagram essentially carry over information introduced in Table I. The left side of the diagram contains new information. It depicts the three skills Speaking (S), Listening (L), and Reading (R) as irregularly shaped forms in definite spatial relationship to each other.

S is portrayed in a shape like a catcher's mitt, L in the shape of a peanut, and R is shaped like a feather. The upper part of the catcher's mitt S encloses the upper part of the peanut L. The lower part of the peanut L impinges on the upper part of the feather R. The lower part of the catcher's mitt S curves toward the lower part of the peanut L and the base of the feather R.

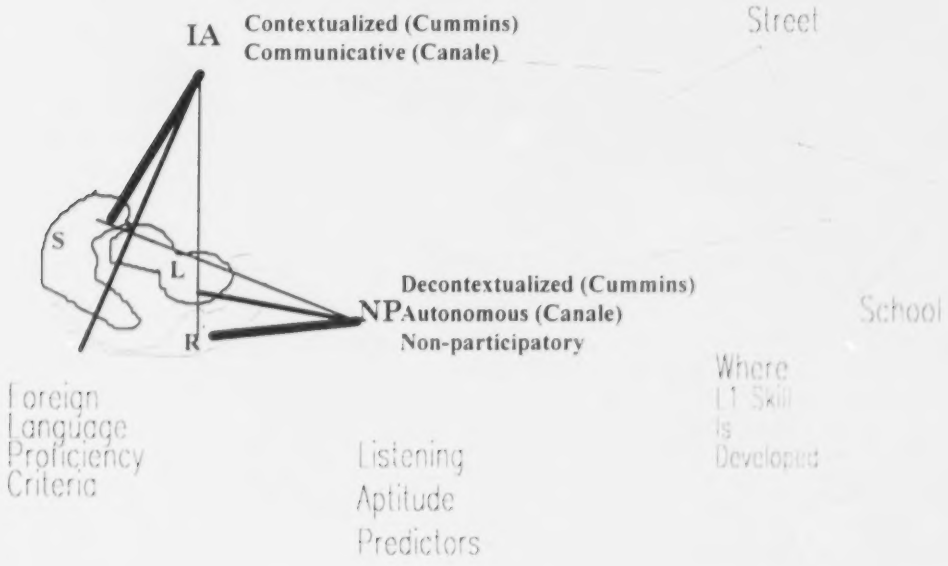


Figure 1

L1 LISTENING AS A PREDICTOR OF L2 PROFICIENCY SKILLS

The following analogies can be drawn. The upper parts of S and L approach each other; this symbolizes the close interaction between S and L in interactive settings. The lower part of L and the upper part of R approach each other; this symbolizes the textual similarities found when listening to formal lectures and reading subject matter texts.

The lower part of the catcher's mitt S approaches the lower part of L and the feather R; this symbolizes planned speech such as lectures. The base of the feather R curls to the right up around toward the back of the catcher's mitt S; this symbolizes the reading of informal notes which bear some stylistic similarity to informal speech.

BEST COPY AVAILABLE

The lines of various thickness from IA (Interactive) and NP (Non-participatory) suggest the possibility that different kinds of listening might have different relationships to L2 skills.[14] [15]

Bottom line: a fuzzy dichotomy of listening. A number of loosely related concepts have been introduced in this section by language analysts writing from different perspectives. Taken together, these concepts suggest the possibility of making a fuzzy dichotomy between different types of listening. The full elaboration of such a dichotomy is beyond the scope of this paper. However, the ideas presented provide a transition to the work of two NL researchers who have analyzed listening into three component traits.

Bostrom and Waldhart's Three Types of Listening

Bostrom and Waldhart (1981) are the authors of the Kentucky Comprehensive Listening Test mentioned earlier. They identified at least three types of listening behavior, which they call short-term listening, interpretive listening, and lecture or long-term listening.

Short-term listening. In the first part of the test, the examinee hears a series of numbers or letters, sometimes accompanied by background noise. He/she is immediately thereafter prompted to answer a question about the order of the numbers or letters in the series. The examinee must respond *immediately* after the prompt. The authors calls this "short-term listening" (STL).

In the second part of the test, the examinee hears again hears a series of numbers or letters, but no background noise. He/she is prompted to answer a question about the order of the numbers or letters in the series *only after an interval* of 20 to 50 seconds after the last number or letter in the series is presented. The authors call this short-term listening with rehearsal" (STL-R).

Interpretive listening. In the third part, the examinee hears successive parts of a dialogue consisting of very brief interchanges. It is apparent from nonverbal audio and situational clues that the speakers sometimes say one thing and mean something else. The examinee must answer questions about the intent of the speakers by choosing from very brief multiple choice options. The authors call this interpretive listening.[16]

---

[14]This diagram should be interpreted with caution. For example, Figure 1 does not account for certain plausible assumptions about early language learning. One such assumption would be that phonological coding ability, grammatical sensitivity, and ability to acquire vocabulary play a major role in early language learning. These predictors might predict globally across skills. This section has only suggested some nonspecific intuitions about what kinds of predictors might be represented by IA and NP. These ideas have not been specified well enough here to try to identify IA and NP with any of the standard reference factors in the mental testing literature. For further discussion on the concept of different variables being important at different stages of language acquisition, see references by Higgs, T., Clifford, R. (1982), Upshur, J, Homburg, T.. (1983), de Jong (1994).

[15] It is interesting to note that multi-method, multi-trait analyses of language skills often find clear trait differences in the case of speaking and reading, but tend to find method and trait confounded in the case of listening. One reason for this kind of confounding might be that an interview method of measuring listening might tap the "S-side" of listening, while a multiple-choice test might tap the "R-side" of listening. Thus Figure 1 might offer some insight into the kind of data found in the reference by Dandonoli, P., Henning G. (1990).

[16]This is a concrete example of a kind of test that might measure a kind of L1 behavior that is closer to "interactive" listening than "noninteractive" listening. However, a much broader sphere of infuence is assigned to interactive listening in Table 1 as a whole, much broader than this one test of "interpretive listening" would measure. Identifying this test with this broad concept of interactive listening would probably go beyond the specific intent of Bostrom and Waldhart.

Long-term listening. In the fourth part of the test, the examinee hears a lecture that is approximately 1500 words in length and must thereafter answer multiple choice questions on the lecture. The examinee is not allowed to take notes. The authors call this lecture listening or long-term listening.[17]

An elaboration of Bostrom and Waldhart introducing the concept of "native authentic listening"

Introducing a concept to elaborate on Bostrom and Waldhart's work. Figure 2 uses visual metaphors to portray relationships between the three listening factors found by Bostrom and Waldhart and another concept I will introduce--"authentic native listening."

Hypothesizing an upper anchor for the ILR listening scale. This new concept itself needs to be elaborated. We need to explain why we as FL researchers have a warrant to use this concept. "Native authentic listening" (NAL) is an extrapolation from a FL learning context to a NL context. We are extrapolating to what a "native listener" would be able to do if he had no need of language instruction. I must consider the construct to be an elaboration on my part because FL researchers like myself devote almost all of our attention to the kind of "authentic listening" that language *learners* with various lesser levels of skill can perform. That is, we don't devote much attention to analyzing, diagnosing, and remediating what native speakers can in some sense already do[18] The term has significance to us not because we observe, think and write a great deal about the concept in our own FL research literature, but because we find it a useful icon for anchoring the end point of the proficiency scale (the theoretical ultimate goal of FL instruction) rather than an object of intensive study in itself.[19]

NAL portrayed as a circle inside a triangle in Figure 2. "Authentic language" (NAL) is represented by a circle inside a triangle. NAL is a set containing "authentic listening" tasks and is located inside the circle. The members of the set are "tasks" and not isolated words and grammatical constructions. This implies that each NAL task consists of a binary relationship involving (1) an authentic NL goal for listening and (2) an accompanying authentic NL text.

Inside the Circle. The members of this set of tasks (defined above as binary relationships) are in different locations inside the circle. These various member tasks are at different distances from the corners of the triangle.

Pure traits portrayed as corners of the triangle. The corners of the triangle represent pure traits (PT) roughly analogous to the traits that Bostrom and Waldhart identified. The metaphor intended here is that various tasks within the circle may require different combinations and weighting of PTs. The combination and weighting of PTs required for individual tasks corresponds to distances from the corners of the triangle.

---

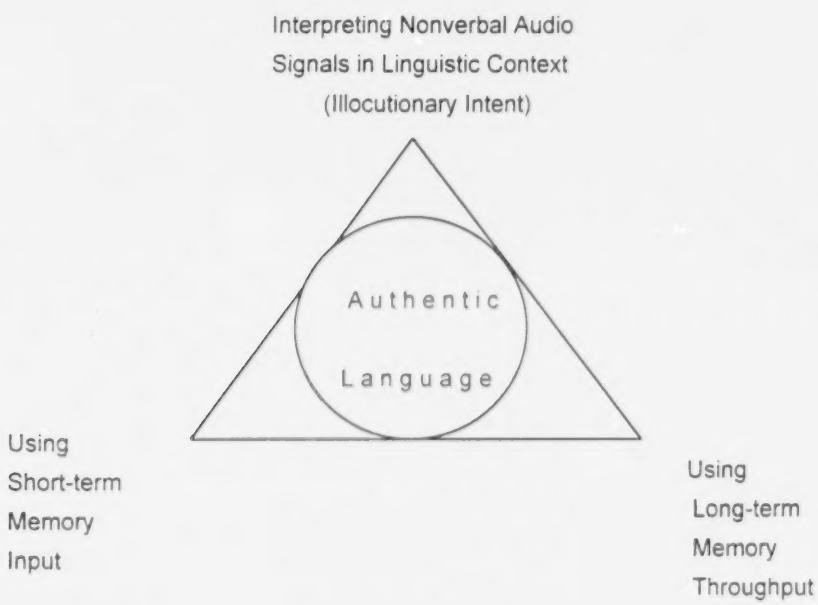
[17]There is also justification for citing this test as an example of "nonparticipatory" listening that is more closely related to reading than speaking.

[18] Consider the following examples of research interests that are seldom found in the FL literature: (1) individual differences in coping ability of native listeners in situations where a speaker introduces new information too quickly for the NL to relate the new information to previously presented ideas, and (2) the kind of notetaking strategies a NL employs in a lecture situation with the intention of later reconstructing and studying the lecture content, in cases where the lecturer presents too many ideas for the NL to follow in real time.

[19]The intention of this extended explanation is to make it easier for those of us in the ILR camp to better communicate with scholars with other research interests by being clearer about our own background interests and thinking.

Why Pure Traits Lie Outside the Circle. The corners of the triangle themselves lie outside the domain of circle that represents NAL. This is a metaphor that has a purpose. It suggests that PTs may predict acquisition of proficiency by language learners without being an NAL task themselves. For example, memory span for letters and numbers is not really a task that belongs to NAL, because NLs seldom make it a listening goal to remember the location of numbers in a string; they don't have any real need to do this as part of their daily life. Nevertheless, memory span for letters *may* predict foreign language proficiency. It is an open question, and an important question, whether a test based on PTs or one based on NAL is a better predictor for the purpose of language aptitude. A broad variety of psychometric and practical issues may bear on the answer to that question.

The scope of these issues is large enough to preclude much discussion of them at this point in this paper. I will return later to the subject of PTs and NAL, and give examples to illustrate the points made above. Before doing that, I want to prepare the ground by addressing yet a third perspective for viewing listening comprehension (in addition to the predictive perspective and the linguistic content perspective). Hopefully, this third perspective will make the examples more cogent.



NATIVE LISTENING

FIGURE 2

Summarizing discussion of predictive perspective and language content perspective.

I conclude this section on the linguistic content approach by expressing another hope. My hope is that the audience perceives there is some connection between one's research background and previous conception of the term "listening" and the number and type of predictors one expects to find under the general rubric "native listening." If that hope is justified, I am ready to present NL from the perspective of cognitive models.

The Last of Three Complementary Approaches in the Literature on Listening Comprehension: the Perspective of Cognitive Models

Introduction

I have two motivations for introducing the topic of cognitive models. One reason is the prominence of the concept in the recent literature. The other reason is more personal. I will start by elaborating my personal interest.

Personal Perspective

I have been struck by the seeming paradox between native listener performance on certain listening tasks invoiving short simple texts and certain other tasks involving long complex texts. In some cases, the native listener will accomplish the task with the long text much more easily than the task with the short text. I will provide a concrete example later. However, I think the example will be easier to understand if I first make use of an analogy to prime the pump. One element in the analogy is the contrast between the native performance on short and long texts. The other element in the analogy involves computer data bases.

If one has a very large data base with a large number of fields, one can create a targeted set of successive queries that quickly selects three or four cases out of 1,000,000 records that have the exact elements desired. On the other hand, if for some reason it is impossible to use an appropriate query, it can be difficult to find a few records in a much smaller data base.

Historical Perspective

Back to the black box. This example about the role of data base queries suggests a path to move from my personalized perspective to a broader perspective. The broader perspective involves the historical development of cognitive models, including models of listening comprehension. There has been a considerable evolution in the past sixty years from the heyday of radical "black box" behaviorism to current day trends in cognitive psychology. A half century ago, many prestigious mathematical psychologists were loosely associated with the behaviorist school. The radical behaviorist school suggested that if we patiently allowed the mathematicians to analyze data on stimulus strength impinging on the black box, response time, and response strength emanating from the black box, their school would eventually explain complex behavior.[20]

There's somebody in my black box. By the 1960s, many prestigious mathematical psychologists had decided to jump ship. These mathematicians had realized that the data about the responses from the black box don't make much sense unless one takes into consideration not only (1) what the organism in the black box must have known before the stimuli came in; but also (2) what the goals of the organism were when it was learning what it now knew; and even beyond that, (3) still more information about what the inside of the organism in the black box must have looked like all the while.[21] Deprived of the prestige mathematicians had contributed to their stimulus-response theories, the radical "black box" behaviorist school no longer had the ability to attract much attention with their own ideas on complex verbal behavior nor to inhibit other ideas from being developed.

---

[20]The progression of thought in the behaviorist school can be traced in the references by Watson (1924); Hull (1943); Skinner (1957). (The classic and decidely antibehaviorist opposing response to the Skinner reference comes from the field of *linguistics;* see reference by Chomsky, N., [1959].)

[21] A continuous process of evolution is evident from the series of references by Hull, C. (1943); Norman, D., (ed.), (1970), Norman, D. and Rumelhart (1970), Greeno, J. (1970), Montague, W. (1977).

Time to talk about different types of memory. Thus, the first evolutionary step occurred when the mathematicians gave a new breed of psychologists permission to hypothesize on what was inside the black box. At first the hypotheses were relatively simple. There had to be a short-term memory, a long-term memory, and some sort of active working memory where a goal-setting executive transformed information from the outside to fit in with previously learned information from long-term memory.

The computer metaphor. The next evolutionary step occurred when individual researchers began to furnish the black box with any additional construct that helped them explain any of their own behavioral data. This was important because technicians in other fields were making progress in fields such as computer data bases, expert systems, and artificial intelligence. All these developments contributed a new source of metaphors to describe the furniture inside an increasingly transparent "black" box.[22]

Introspection returns to favor. The final evolutionary step occurred after introspective (and retrospective) techniques such as think-alouds returned into favor and became familiar instruments in the cognitive psychologists' tool box. Nowadays investigators commonly use language borrowed from the field of data processing to both describe and elaborate introspective and retrospective data.[23][24]

This last evolutionary step provides the context for me to return to my personal concern with tasks involving short and long listening texts. I will now provide concrete examples to use in a think-aloud. I intend to then use the retrospective data from the think-aloud to construct an analogy with a multimedia database.

Concrete Examples

The first example on the following page is a listening task with a short amount of audio text.

The second example is a listening task with a large amount of audio text. The tiny subscript numbers serve only to identify the sentences in the text for subsequent discussion.

Preliminary discussion of the two passages

The two passages are printed on the following page.

A seeming paradox. Small scale trials indicate that native speakers find the second task easier to perform than the first task. Yet the text for the second task is much longer. It also has a variety of features that might confuse a foreign language learner with little proficiency--such as idioms, reasonably complex grammar, and somewhat culture specific content. If we remember the metaphor of the vast ocean and the vast variance, we might suspect that there are some interesting things to consider about the second passage. These interesting things could correspond to the residue left after our hypothetical ocean evaporated.

---

[22]The progressive development of computer analogies can be traced in the references by Anderson, J., Bowen, G. (1974), Findler, N., (1979), (ed.), Cermak, L. Craik F., (1979), Kolodner, J. (1984). In turn, these computer analogies were intellectually compatible with development of "spreading activation" and related "connectionist" theories in references by Collins, A., Loftus E. (1975); and Cottrell, G., (1994).

[23]The reference by Hintzman (1987) provides a balanced historical overview of the competition and interaction between behaviorist and cognitivist schools of psychology., and summarizes in accessible form the path of evolution represented by the references in footnotes 16-18.

[24]See reference by Faerch, C., Kasper, G. (1987) concerning use of introspective techniques in second language research. Think-aloud techniques also play a role in the language learning strategies literature. See references by Wenden, A., Rubin J., (ed.), (1985),Thain, J., Lett, J., (1991).

AUDIO EXAMPLE 1

Listen to the following series of numbers and be ready to answer a question about the numbers:

024
252
306
408
503

What was the third number presented?

AUDIO EXAMPLE 2

Listen to the following text and be ready to answer a question about the text.

[1]Let's look at my schedule before I give you a number to call and tell you when you should call about your file. [2]If I'm in a meeting or in someone else's office, people will be around nipping at my heels, you see, and not only that I may not have my stuff at hand to talk to you.

[3]From 9 to noon, everybody including me too, will be at extension 463, that is in theory, but we'll all be behind closed doors at the contract award board.

[4]From 1 to 3, I'll sneak back to my files at my old office at extension 654, where every one is on leave anyway. [5]1 to 3 at 654, make a note.

[6]From 4 to 6, we'll all be back at the Contract Approval Office at extension 625, all of us huddling together to tie up all the loose ends from the morning again, so if the unexpected happens and you can't get me earlier in the afternoon, this is a last resort to call 625 then.

When should you start trying to call me and at what number?

The need to focus. In the first task, the listener probably attempts to hold previous linguistic input from the speaker in his memory in its *original* form, while the speaker continues to provide new input. In this task, the listener (L) *might want* to identify which input is more important and which less important, but the structure of the task gives L *no* opportunity to do so. If L only had a goal that enabled L to decide which input deserved more attention, L might be able to make the important input *more* salient in L's own mind than any less important input that might come later. Unfortunately, L has no clue as how to accomplish this, and thus has no way of preventing later and less important information from driving what ultimately turns out to be important information from L's working memory. If L were to give a list of appropriate verbals and verbal combinations that describe what L would *like to do, but can't do* in this task, that list might include such words as rehearsing, activating/maintaining, focusing, and attending.

Having a goal helps. In the second task, the listener will probably quickly give up on holding most of the input in its original form. Instead L quickly realizes the task is structured in such a way that L can almost immediately define a goal and begin to assimilate important information into larger cognitive structures. The cognitive structures will comprise an interlocking set of interpretive schemata. The seed template for the larger structures existed in some sense in L's long-term memory before the listening task began. Such templates were based on the L's broad past experience.

AUDIO EXAMPLE 2

Listen to the following text and be ready to answer a question about the text.

[1]Let's look at my schedule before I give you a number to call and tell you when you should call about your file. [2]If I'm in a meeting or in someone else's office, people will be around nipping at my heels, you see, and not only that I may not have my stuff at hand to talk to you.

[3]From 9 to noon, everybody including me too, will be at extension 463, that is in theory, but we'll all be behind closed doors at the contract award board.

[4]From 1 to 3, I'll sneak back to my files at my old office at extension 654, where every one is on leave anyway. [5]1 to 3 at 654, make a note.

[6]From 4 to 6, we'll all be back at the Contract Approval Office at extension 625, all of us huddling together to tie up all the loose ends from the morning again, so if the unexpected happens and you can't get me earlier in the afternoon, this is a last resort to call 625 then.

When should you start trying to call me and at what number?

I eventually want to retrace my steps in the previous paragraph, and illustrate why a multimedia data base is a good analogy of the process I am describing. However, I will first prime the pump by briefly elaborating on the function of the larger cognitive structures mentioned in the previous paragraph.

Activating important information and forgetting the rest. The larger cognitive structures will accomplish more than merely assimilating the original information. They will also (1) assimilate succeeding pieces of information that are important in terms of the goal, (2) keep the important information active in working memory, (3) deactivate less important information. Furthermore, the effort required to keep the larger cognitive structure alive in working memory will place less load on the listener's cognitive resources than would a corresponding effort to preserve isolated pieces of information in memory. The new structure will help the listener (1) fill in the gaps beyond what the speaker has explicitly said, *and* (2) "edit out" (into an inactive state) some unimportant things that the speaker actually did say. If L were to give a list of appropriate verbals that describe what L is able to accomplish in this task, that list might include such words as elaborating, interpreting, activating/absorbing, and inferencing.

The Active Listener and the Analogy of a Multimedia Database

Now I can retrace my steps and address the question of why a multimedia data base is a good analogy for what is happening in the second task.

(1) Upon hearing the first sentence in the text "Let's look at my schedule before I give you a number to call and tell you when you should call about your file.", L consults the "data base" under a field named GOALS, and finds a template that matches the input. This template probably tells L to be ready to conduct another search based on fields such as TIME, LOCATION, PHONE EXTENSIONS, FILES, and SPEAKER GOAL to match the expected input.

(2) Upon hearing the second sentence L suspects L should be ready to take any further input and conduct a major sort on PERSONS and a minor sort on LOCATION and PHONE NUMBER, with two intentions in mind. The first intention is to deactivate any piece of incoming information in which more than one PERSON is present. The other intention is to concentrate on any record in which the speaker is the PERSON. In addition, L infers that L should be ready to take the LOCATION and PHONE NUMBER fields of the remaining records and be ready to run major sorts on these fields with minor sorts on SPEAKER INTENT, TIME, and FILES.

AUDIO EXAMPLE 2

Listen to the following text and be ready to answer a question about the text.

Let's look at my schedule before I give you a number to call and tell you when you should call about your file. If I'm in a meeting or in someone else's office, people will be around nipping at my heels, you see, and not only that I may not have my stuff at hand to talk to you.

From 9 to noon, everybody including me too, will be at extension 463, that is in theory, but we'll all be behind closed doors at the contract award board.

From 1 to 3, I'll sneak back to my files at my old office at extension 654, where every one is on leave anyway. 1 to 3 at 654, make a note.

From 4 to 6, we'll all be back at the Contract Approval Office at extension 625, all of us huddling together to tie up all the loose ends from the morning again, so if the unexpected happens and you can't get me earlier in the afternoon, this is a last resort to call 625 then.

When should you start trying to call me and at what number?

(3) Upon hearing the third sentence, L carries out the planned queries, and deactivates the information because the PERSONS field does not match.

(4) Upon hearing the fourth sentence, L carries out the planned queries again, and saves LOCATION, PHONE NUMBER, TIME, and FILES from the input and still has resources left to check the input against SPEAKER INTENT.

(5) Upon hearing the fifth sentence, L verifies SPEAKER INTENT, and activates the follow record: LOCATION (my old office), PHONE NUMBER (654), TIME (1 to 3), FILES (Present), SPEAKER INTENT (Helpful toward meeting listener goal), and GOAL (know where and when to call about file). L will now check any incoming information against this record and deactivate any nonmatching record.

(6) Upon hearing the sixth sentence, L is ready to deactivate incoming information to prevent interference with the previously validated record. This is because the information in the sixth sentence doesn't match all the fields in the previously validated record, (e.g. LOCATION (Contract Award Office), FILES(Inferred to be absent), SPEAKER INTENT (busy solving another problem). By this time L could have forgotten the first phone number because L had already deactivated it. L is also ready to place a priority on rehearsing the record with TIME(1 to 3) and PHONE NUMBER (654), with secondary priority on remembering the last PHONE NUMBER (625), which matches only on SPEAKER INTENT (gives number as last resort).

(7) At this point the test question is given. As soon as L verifies that the activated record is the answer to the question, L fine tunes the GOAL to (provide answer), provides the answer, and deactivates all other information.

(8) There is another field in L's data base that will be activated during this conversation. However, the input matching against this field cannot be localized to a single sentence. If one omits the words and simply hums the discourse intonation, one finds that the intonation itself gives a strong indication where the most important information is.

<u>Lessons to be Learned from these Two Passages</u>

Before proceeding, I will summarize what we can learn from the two passage examples:

Try it, you'll learn something. First of all, I concede to skeptics who think I stacked the deck with these examples to make a rhetorical points that they are right, and I will proceed to make those very rhetorical points. However, I do suggest that interested readers attempt small-scale experiments like the one described above to convince themselves from their own experience that the variables described do play an important role in NL.

Useful database analogy. The data base analogy has been helpful in illustrating that features other than vocabulary, grammar, and passage length can affect NL comprehension. On the other hand, a nonnative speaker with a lesser level of proficiency might have been distracted by some of the very parts of the passage that helped the NL perform the task.

Pedigree of database analogy. I chose to use the analogy of using a database to show how a listener might elect to select certain information and ignore other information. Those familiar with other connectionist approaches might correctly think that my informal analogy has some parallels with these approaches. In brief, a connectionist approach suggests that all the various elements (words, inferred pragmatic goals, grammar, intonation) at *different* levels of linguistic (and perhaps some metalinguistic and nonlinguistic) structure in the spoken input are involved in interpreting an incoming message.[25] They are involved in the sense that they all get to "vote" on what kind of interpretations make sense in terms of the intent of the incoming message. Interpretations that are "voted" as plausible are activated and implausible interpretations are ignored. Activated interpretations provide the context for interpreting the input that follows. Certain elements are more likely to be "connected" or "associated" with each other by context. One can visualize a number of different "images" to represent this kind of "connection:"

(a) In my data base analogy a series of queries scored "hits" or "matches" that influenced successive searches.

(b) Another image might be that "connections" that are stronger support each other (vote for each other) in context and "veto" other less plausible connections.

(c) Another image might be that "connections" that are inherently more plausible in context are awarded more votes and outvote other possibilities.

Forerunners of contemporary connectionist approaches include Collins and Loftus' (1975) theory of spreading activation and Anderson's (1983) adaptive control of thought (ACT).

Recent applications of similar models in artificial intelligence have succeeding in producing machines that can carry on a surprisingly natural conversation within certain limited topic domains. This success seems striking enough to lead me to speculate further on the kind of cognitive abilities required for comprehension skills.

Not just a database, but a multimedia database. I have suggested an analogy be made between the listening process and a *multimedia* database--not just an ordinary database. In order to make this analogy clear, I will elaborate on some of the characteristics of a multimedia database. In a multimedia database, elements might be in text form for some fields, but in the form of video or audio for other fields. The user of such a database might have the capability to inspect the text fields and at the same time call upon peripheral devices to view or listen to the audio and video elements in other fields. This suggests an analogy to the listening process.

---

[25]The metaphor that "a listener actively uses a database to process ongoing discourse" is also compatible with the assumption that the NL tacitly assumes and proactively employs Grice's (1975) maxims to help infer linguistic and discourse structure at every linguistic level, especially the pragmatic level.

The analogy would involve mental processes during listening in which "nontext" elements such as (1) voice affect and (2) intonation patterns could be grouped together under "fields" to be searched. The NL would conduct queries in order to choose matching interpretive schemata to focus his/her ongoing listening process and to deactivate irrelevant schemata during subsequent listening. Just as real mechanical peripheral devices have performance limitations that can be objectively studied, I would hope that connectionist mental models would provide a basis for studying the characteristics and limitations of mental subsystems contributing to listening comprehension. In addition, the mental measurements specialist may find connectionist models suggest hypotheses as to what measures are appropriate to predict and measure comprehension ability and language acquisition. For example, they might speculate that measures testing the processing of vocal elements less directly involved in lexical processing may provide sources of variance distinct from those measures typically associated with strictly lexical processing. This analogy thus suggests the possibility that NL testing should use two distinct listening measures: a "lexical focus" listening measure and a "voice focus" listening measure.[26]

Those that have nothing to seek take longer to find. Real-life database users know well the frustration caused when they try to find a certain single record in a large database file, but don't have a clear idea of what query to use. Sometimes they have to just give up and turn their attention to more pressing business. This familiar experience from the computer world may have a parallel in listening comprehension. Spearitt (1962) administered a large number of listening comprehension measures along with other cognitive tests. He found that tape-recorded tests with such names as Illogical Grouping and Haphazard Speech loaded on a memory span factor.[27]

Spearitt's findings tie in with several other ideas presented in this paper. After our experiment with the short text and the long text, I suggested that the presentation of the shorter text did not allow the listener the opportunity to establish a goal in time to chunk the important input into a larger cognitive structure. It is reasonable to suppose that a longer memory span would give a listener a little more time to hold input in short-term memory before deciding how to chunk it into an appropriate structure.[28]

The argument in the preceding paragraph suggests that we can add a "memory span" variable to the "lexical focus" and "voice focus" variables mentioned above. This is a conclusion similar to the one Bostrom and Waldhart reached through a different route, when they established a distinction between short-term listening, interpretive listening, and long-term (lecture) listening. Of the three traits, only long-term listening seems to have something in common with the measures presently included in the ASVAB and DLAB at this time.

Conclusions and recommendations concerning NL measures

Our review of NL tests and of the literature on listening has enabled us to come to some tentative conclusions. However, since we at DLI don't have much experience in actually writing NL tests. We would like to seek out the opinions and help of experts who have had more practical experience. For this

---

[26] References by Doff, A., Jones C. (1980) and Haycraft, B., Lee, W. (1982) are basic ESL conversational course materials, but with a special twist that may give the reader a hint of some of the kind of skills might be involved in "voice focus" listening measures.

[27] See Carroll's (1993) reanalysis of Spearitt's data set one one of the series of diskettes accompanying Carroll's recent book cited in this reference.

[28] A variety of other studies have addressed a number of relationships between memory span, speed of auditory closure, listening to distorted or illogical speech, and listening to speech with background distractions. The results of these studies seem to be influenced by the variety in testing measures employed and by the specific populations chosen. See references by Karlin, J. (1942), Stankov, L., Horn, J., (1980), Horn, J., Stankov, L. (1982), and related comments by Carroll, J. (1993).

reason, as I present each tentative conclusion, I will also identify areas in which we at DLIFLC might benefit from the expertise of other scholars.

Decision criteria for evaluating alternatives

A good starting point is to list the criteria for evaluating alternative L1 listening test types for inclusion in an expanded language aptitude battery. It is hard to improve on Henning's (1987) largely self-explanatory list of criteria for evaluating language tests, which I quote below:

Purpose of the test:	test validity
Characteristics of examinees:	test difficulty
Precision and accuracy:	test reliability
Suitability of format and features:	test applicability
The developmental sample:	test relevance
Availability of equivalent or equated forms:	test replicability
Scoring and reporting:	test interpretability
Cost of test procurement administration and scoring:	test economy
Procurement of the test:	test availability
Political considerations:	test acceptability

I might add two other criteria relevant to our plans to expand the current DLAB: (1) since tests to be retained from the old DLAB already require 75 minutes to administer, it is undesirable for the total test administration time of an expanded DLAB should exceed two hours; and (2) in order to use the total administration time wisely, DLI would like to avoid adding measures that duplicate any part of the current DLAB or ASVAB, the two screening batteries used to select students.

The above list of criteria gives an idea of DLI concerns. A complete evaluation of NL tests in terms of all these criteria is far beyond the scope of this paper.

Pure traits (PTs) vs. Native Authentic Listening (NAL): A Quick Scan of Current NL Tests as Models

I made a distinction earlier between measures of PTs (pure traits) and NAL (Native Authentic Listening). The concept of measuring PTs derives from a tradition in mental measurements that places a high value on defining minimally intercorrelated traits, --sometimes even at the seeming expense of ecological or face validity in test content. I noted that FL methodologists see NAL as a *theoretical* ideal (in terms of face validity), because they can equate NAL with the upper anchoring point for the ILR proficiency scale for *FL* listeners. The diagram presented earlier in Figure 2 and the accompanying explanatory text explained the relationship between PTs and NAL.

I left the question open as to whether PTs or NAL were the most appropriate measures of NL comprehension as a predictor of L2 proficiency.

Table 2 attempts to list the *number* (expressed by a digit in large type) of instances in which each of the reviewed NL tests contain (1) item types that measure PTs, (2) item types that measure NAL, and (3) item types that are on the borderline between PTs and NAL. The table serves to roughly quantify the occurrence of these types of items in these tests. It is hazardous to draw detailed conclusions from this table, because it does not furnish a very precise categorization of item types. The main conclusion that can

TABLE 2
NL LISTENING TEST ITEMS: PTs or NAL?

Name of Test	Number of distinct item types in each test that tend to measure PTs rather than NAL	Number of distinct item types in each test that measure on the border of NAL, and thus tend somewhat toward measurement of PTs	Number of distinct items types in each test that clearly measure NAL, not PTs
Watson-Barker		2 !Lecture listening !Emotive listening	3 !Conversations !Instructions/Directions !Listening for Content
Kentucky Comprehensive Listening Test	1 !Short/-term memory (2 types)	2 !Lecture listening !Interpretive listening	
Carleton University Test	1 !Lecture listening as bootstrap to library research		
NTE Communicative Skills Listening Test		2 !Interactive situations involving empathic listening !Lecture listening to extended passages on educational topics	1 !Variety of listening situations especially school situations without strong cognitive or emotional load
NTE School Guidance and Counseling	1 !Counseling situations involving empathic listening		
Brown-Carlsen	1 !Immediate recall	1 !Lecture listening	1 !Miscellaneous other item types
STEP	1 !Immediate recall	1 !Lecture listening	1 !Miscellaneous other item types

be drawn[29] from Table 2 is that some of the item types measured on NLs are more like measurements of PTs than NAL[30], some of the item types clearly measure NAL, and some are on the border line (the edge of the circle in Figure 1.)[31] Thus, a survey of NL test item types does not in itself give any guidance as to whether one should proceed with a PT measurement approach, an NAL approach, or something in between.

The next two sections deal with the kinds of considerations involved in using PT test content and NAL test content in NL tests used as aptitude tests for predicting FL proficiency.

Measures of PTs as FL aptitude test measures

Three PTs were identified earlier in Figure 1. They involved short-term memory, long-term memory, and interpretation of nonverbal audio signals.

Relation of PT measures to current and future ASVAB. PT measures of long-term memory may tend to share some variance with ASVAB tests that are associated with cognitive and verbal achievement. Furthermore, although there is no short term memory test on the current ASVAB, working memory tests that tap similar abilities have been proposed for inclusion in ASVAB. On the other hand, nothing in current or projected ASVAB versions will test nonverbal audio signals.

Nothing in the current DLAB seems to compare to any of the three PTs.

Using PTs measures of long-term memory and lecture listening measures: choice of content areas. Performance on long-term or lecture listening tasks is facilitated when a listener has access to content-area schemata for the subject areas represented in the listening texts. Depending on the circumstances, knowledge of almost any content area schema acquired prior to L2 study could potentially be useful in L2 listening, especially after the L2 listener has surmounted initial phonological, grammatical, and lexical hurdles.

However, it is likely that some broadly conceived content-area schemata would be particularly relevant:(1) international and cross-cultural communication; (2) issues of sensitivity to international and cross-cultural differences; (3) international business, political, cultural, and military cooperation (or rivalry), (4) cross-cultural technological transfer (or maintenance of technological secrecy); and (5) comparative political science. On the other hand, one could easily name a number of content area schemata that

---

[29] My sources of information were test information brochures, published information, and personal communications, rather than a detailed review of the physical contents of each test. In some cases, I have combined what the publisher considered two or more item types into a single item type to more simply fit into my classification scheme.

[30] (The reader may wish to simultaneously refer to Table 2 below and to Figure 2 [which was presented earlier] to follow this footnote.) I identified short-term memory tasks and immediate recall tasks as PT measurements. They fall outside NAL near the "short-term memory" corner of the triangle. I identified the Carleton University task with PT measurement, since general academic ability is important in carrying out that task. This test falls outside NAL near the "long-term memory throughput" corner of the triangle. Similarly, the kind of listening in the "School Guidance and Counseling Examination" is located near the "illocutionary interpretation of non-verbal signals" corner of the triangle.

[31] In cases, where the trait specialization is not as striking as in the previous footnote, I locate lecture listening on the border of NAL and tending toward "long term-memory," whereas I place "emotive listening" and "interpretive listening" on the border of NAL and tending toward the "interpretation of nonverbal illocutionary intent" corner of the triangle.

(especially if narrowly interpreted) would probably be less useful in learning to listen in a language--schemata for abstract mathematical concepts, American sports history, local American building codes, and personal histories of American radio and television entertainers, come to mind as examples.

This suggests that there is a tradeoff to consider in the choice of topic areas for lecture listening tasks. The broad nature of potential applications of foreign languages implies that the choice of topics should be relatively general, but the very nature of career interests of FL listeners suggests that some topics are more appropriate than others. This issue is not only salient for the task for designing a listening component for a FL aptitude test. The publishers of NL tests generally have an "*occupational content target area*" to guide their selection of content; they probably also have to think about striking a balance between general and specialized topic areas. This is one area in which DLI could probably learn from an exchange of experiences with the writers of the NTE Basic Communication Skills Listening Test, the Watson-Barker Listening Comprehension Test, the Kentucky Comprehensive Listening Test, and the test used by Carleton University.

Using PTs as measure of listening skills involving perception of affect. Just as *content* schemata help the listener understand lectures, it is likely that *situational* schemata help the listener understand audio messages with strong affective overtones. It makes sense to talk about a *situational target area(s)* for a NL listening test including an measure of sensitivity to affect. As in the case of occupational content target areas, the situational target area(s) for a NL listening test used for aptitude prediction could differ from target area(s) of such current NL tests as the NTE Basic Communication Skills Test, the NTE School Guidance and Counseling Examination, the Kentucky Comprehensive Listening Test, or the Watson-Barker Comprehension Test. All of the above tests vary in the number of individual test items, the number of situations, breadth of coverage across situations, item length, degree of context provided, and the extent to which cognitive information and affective information are both presented in the same text.

One concern is that the danger of subjectivity or low reliability in tests that measure mainly sensitivity to illocutionary intent or affect, rather than objective cognitive or semantic information.[32] On the other hand, if tests that measure only affect could be made reliable, these tests could turn out to be a potential new source of variance and predictive power. This is because these tests may not share much covariance with verbal and mathematical factors on ASVAB, or with phonological coding and grammatical sensitivity factors on the current DLAB.

The authors of the NL tests mentioned above had to consider a balance between (1) general and specialized situations; (2) long and short items; (3) items involving cognitive knowledge *and* situational sensitivity as opposed to items in which only situational sensitivity seems to matter; (4) and between alternatives in overall content coverage in test planning. The content coverage in some sense has in each case to be appropriate to the career interests of the potential test examinees and the purposes of the test. Again, this is area in which DLI could probably learn by exchanging experiences with NL testers as to how to select test content appropriate to the career focus of the FL linguist.

Measures of NAL as FL aptitude test measures

It is possible to base discussion of test content solely on NAL, rather than PTs. However, even if all the test content was genuine NAL, one could still suppose that each component item would represent a task that requires some cognitive contribution from each of the PTs. Some item tasks would require greater cognitive contributions from some PTs than other PTs.

For example, NL for certain kinds of instructions and directions could place more of a load on short-term memory than long-term memory or illocutionary sensitivity. Certain other NAL items could easily involve NL tasks that place higher demands on either: (1) affective and situational sensitivity, or (2) cognitive or academic sensitivity.

---

[32]See Bostrom, R. (1990a), p. 19.

Some of the same issues in content selection mentioned above in the discussion of PTs would also thus apply even for a test focused on NAL. From this point of view, DLI could benefit from exchanging experiences with the writers of tests like the NTE Core Battery Listening Test or the Watson-Barker Listening Test. These tests have placed somewhat less emphasis on breaking NL into separate or specialized traits than have some of the other tests listed above.

Bottom Line on NL tests as predictors

Several kinds of L1 listening tests are likely candidates for an FL aptitude battery.

Since DLI doesn't have much experience in actually writing NL tests, our agency could benefit from interaction with NL researchers with interests outside the FL testing field. There has not been a great deal of communication of between the disciplines of FLL and NL research. This is an area where DLI could foster a basic exchange of information concerning research interests and backgrounds between researchers in these two disciplines. Subsequent interdisciplinary exploratory efforts could play a very important role in the revision of the DLAB.

Subject to feedback resulting from such interdisciplinary interactions, I can draw certain tentative conclusions.

General conclusions. It would be best if the addition of a L1 test should not greatly increase the length of the DLAB. A revised DLAB (including both old retained tests and new added tests) should not exceed two hours in administration time. Optimal administration time would be somewhat less than two hours

There should be no copyright or licensing problems that would prevent unrestricted duplication and subsequent administration of tests by the Department of Defense (DoD). DoD would want to retain unfettered controls over the administration and test security of any test added to the DLAB.

As explained earlier in the section on the predictive perspective, DLI should consider adding tests that are different from any test currently used in ASVAB and DLAB, the currently used screening instruments. The rationale for having a different kind of test, is that a different test is more likely to measure something new and not duplicate variance already measured by another test.

Test content. From this point of view, DLI should consider tests of short-term memory and tests focusing on vocal quality or sensitivity to illocutionary intent. These tests might be less likely to duplicate the verbal factor variance found in ASVAB. Tests of such abilities might be designed in such a way to also measure auditory perceptual closure and resistance to distraction and auditory distortion. Alternatively, one could consider separate tests for perceptual closure and resistance to auditory distraction or distortion. Although it is desirable that new abilities be measured, DLI needs to also be concerned with the reliability of potential new measures. Of course, it is doubtful that an unreliable test can contribute much additional predictive power to a revised battery.

DLI should not completely exclude the possibility of adding listening tests that are likely to load on a verbal factor. If we elect to design such tests for inclusion in DLAB, we should consider focusing on occupational and situational content target areas. However, one should also consider including a broad range of content areas corresponding to great number of potential applications of foreign language proficiency.

Exploring Tests of Grammatical Sensitivity in English

A review of the tests of grammatical sensitivity is presented. It is not accompanied by an extensive review of the literature on testing grammatical skills comparable to the review of the NL literature given earlier.

The review comprises both: (1) tests of sensitivity to English grammar, and (2) tests that measure sensitivity to foreign (or artificial) language rules.

In contrast to NL tests (which have never before appeared as parts of an language aptitude battery), some tests of grammatical sensitivity have previously been incorporated as parts of aptitude test batteries.

Tests of sensitivity to English grammar

English Grammar Recognition Test (EGRT)

The EGRT was developed at DLI in 1975. It measures explicit knowledge of grammatical terminology. An example of the type of item found in the EGRT is given below:

> A word that modifies a verb or adjective by expressing time, place, manner or degree is called:
>
> a. intensifier
> b. gerund
> c. adjective
> d. adverb

The Flanagan Expression Test (FET)

The Flanagan Expression Test, published by Science Associates, does not require knowledge of grammatical terminology. It has two parts.

In Part One the examinee must identify whether each of a series of English sentences is correct in terms of grammar or usage. An example of a Part I item is given below:

R W I done the work at home.

In Part Two, the examinee must identify which one of three sentences is the "best" way to express an idea.

> ___Most of Greenland consists of glaciers and barren highlands, and no more than two per cent of the island is inhabited and so it is very sparsely populated.
>
> ___Greenland is very sparsely populated. Barely two-percent of the island is inhabited, the rest consisting of glaciers and barren highlands.

The test as a whole has 50 items and takes a little over five minutes to administer. Thus the test is heavily speeded.

DLI efforts to conduct statistical analysis on the FET have been hindered by the fact that student responses to the FET must be recorded on a proprietary non-machine scorable answer sheet.

Preliminary analyses suggest that a large part of the test variance in our test population might be accounted for by a small number of items measuring case and number agreement.[33]

MLAT Part IV (Words in Sentences)

The Modern Language Aptitude Test (MLAT) is published by the Psychological Corporation. It has five parts. Part IV is designed to measure ability to understand the function of words and phrases in sentence structure, without calling upon knowledge of grammatical terminology. Each item consists of a key sentence with a word or phrase printed in capital letters, followed by one or more sentences with words or phrases underlined and numbered. The examinee is directed to pick the word or phrase in the second sentence or sentence group which does the same thing in that sentence as the capitalized word does in the key sentence. An example of the type of item found in MLAT Part IV is given below.

He spoke VERY well of you.

Suddenly the music became quite loud.
1 2 3 4

Tests that measure sensitivity to foreign (or artificial) language rules.

Pimsleur Part IV (Language Analysis)

The test booklet presents a number of words and sentences in Karbardian (a language spoken in the former Soviet Union), and their English equivalents. From these examples, the examinee must figure out how to say 15 new sentences in Karbardian. The items require the application of the examinee's sensitivity to grammatical systems. The examinee is given twelve minutes to answer 15 items.

DLAB Part III (Foreign Language Grammar)

The examinee's task is to learn some grammar rules of an artificial language and then apply these rules in the translation of short phrases and sentences. The words and sentences of the artificial language are similar in some respects to those of English in pronunciation and meaning but have been transformed by the application of rules of the artificial language morphology and grammar. For each item in the test, (1) the examinee *reads* an English phrase or sentence in the booklet, (2) *listens* to the four *alternative translations* in an artificial language spoken on the test *audiotape*, (3) and marks the correct translation on the answer sheet.

The test is so designed that the examinee is effectively discouraged from using a consistent strategy of "reasoning out" the rules to produce a correct answer. For example, (1) the English sentences to be translated are on a separate page from the rules; (2) the examinee is mentally focused on listening to the audio multiple-choice options on the tape; and (3) the examinee cannot review all the options at the same time because the options are presented in serial order on the test audiotape.

Thus as the test progresses and increasingly more grammar rules are introduced, the examinee must become progressively more dependent on automatic processing of previously presented grammar rules.

---

[33] In all of these items, the noun phrases that govern the agreement include either coordinate or complex noun phrases. Strong individual response differences are found in Part I items with stimulus sentences of the type "The videotape playback shows that each of the men and women notice the thief breaking in the office." It is unclear whether individual student differences in answering these items arise from failure to understand a grammatical rule or its scope of application, or from difficulty in applying the rule due to a combination of test speededness and grammatical complexity of the governing construction.

DLAB Part IV (Foreign Language Concept Formation)

The examinee sees four pictures at the *top* of every page of the test booklet. Each picture is accompanied by a description in *an artificial language* of the object or activity depicted in the picture. Taken together these associated pictures with artificial language text constitute a "linguistic corpus" at the top of the page that an examinee must utilize to find correct answers to test items printed on the *bottom* half of that page.

Each item in the bottom half of the page consists of (1) a *picture* and (2) *four written multiple choice options* in the artificial language.

The examinee must find appropriate analogies based on the information in the corpus *and* the individual item to determine which *option* should be matched with the numbered *picture* for that item. The test is moderately speeded.

A completely different set of pictures in a completely different artificial language is introduced on *each* succeeding page. In order to complete the analogies on each page, the examinee has to determine what type of information is relevant to solve the problems on that page. The needed information might be the main concepts underlying each set of pictures, or the graphemic, morphological, or syntactic similarities between the corpus and the individual options for each item. Thus the examinee must have a sensitivity for what kinds of grammatical, morphological, and semantic analogies are possible in a foreign language to solve the problems represented by each item.

Bottom Line on Grammar Tests as Predictors

General

I have completed a review of some tests of grammatical sensitivity, but I have not yet gone ahead to review the literature and issues related to the use of such tests as language tests.

The review of NL tests and literature might provide a useful model for a follow-up review of grammar tests. As in the case of NL tests, DLI could address the utility of grammar as aptitude tests from three perspectives. I will sketch a tentative idea of the components of such a three-part review below.

Predictive Approach

It would be important for DLI to consider the predictive perspective for grammar tests in much the same I did for NL tests. The goal would be to identify the kind of grammar tests that would be most likely to add another source of predictable variance, and less likely to duplicate variance already measured in the current DoD linguist screening process.

"A Grammar Learning Factor Approach"

The next approach in the review of the NL literature was the linguistic content approach. However, grammatical sensitivity is not itself one of the four language skills, but a factor that cuts across all of the four skills. Furthermore, there has been considerable evolution in thinking and ongoing debate for many years as to the proper role of grammar in language learning, and especially to the contribution and relevance of grammar to language learning at various points of the ILR scale. In a review of *grammar* tests, the second approach might be better named the "*grammar learning factor approach.*"

A section devoted to this approach might identify different skills measured in tests such as the EGRT (knowledge of grammatical terminology and ability to apply such terminology in formally analyzing sentences), MLAT (ability to detect parallel grammatical functions and structures in pairs of sentences), and the Flanagan Expression Test (ability to identify grammatical and stylistic correctness under speeded

conditions). It would be profitable to investigate how foreign language methodologists rate each of these tests in terms of "face validity." Such ratings would no doubt influenced by their own backgrounds in teaching foreign languages and analyzing foreign language acquisition. Such backgrounds, however relevant to foreign language instructional experience, might need to be supplemented by information from a third perspective.

A Cognitive Models Approach

The last approach in the review of NL literature was the cognitive models approach. A parallel approach devoted to grammar tests might focus on experimental psycholinguistic research and studies of computational parsers. Psycholinguistic research of this type might be concerned with human parsing preferences where multiple grammatical clues are present. This type of approach might lead in different directions from the second approach. The second approach, as suggested above, is grounded in classroom language teaching experience rather than formal analysis of the operation of grammatical systems.

Where We Go from Here

Although I have not conducted an exhaustive literature review, I am certain there is an abundance of literature corresponding to each of the three approaches, but no concise synthesis of how the three approaches might relate to the use of grammar tests as language aptitude measures.

I think an intermediate step is needed before DLI develops such a synthesis on its own. DLI should continue to foster an exchange of ideas about the role of grammar in language acquisition and about the role of grammar tests in language aptitude testing. Scholars in the fields of foreign language methodology, psycholinguistics, and cognitive psychology could make valuable contributions to this exchange. Hopefully, these contributions would be a stimulus for DLI to conduct a thoughtful review of the literature at a later time. The intent of this review of would be to evaluate specific types of grammar tests for inclusion in a revised DLAB.

References

Anderson, J. (1983). *The architecture of cognition.* Cambridge: Harvard University Press.

Anderson, J., Bowen, G. (1974). *Human associative memory.* Washington: Hemisphere Publishers.

Arnold, C. (1982). On listening: what does rhetoric have to say to cognitive psychology? In Weimer, W., Palermo D., *Cognition and the Symbolic Process, Vol. 2,* 131-157. Hillsdale, NJ: Lawrence Erlbaum.

Bostrom, R. (1990a). Measuring individual differences in listening. In Bostrom, R., (ed.), (1990) *Listening behavior: measurement and application.* New York: Guilford Press.

Bostrom, R., (ed.), (1990b). *Listening behavior: measurement and application.* New York: Guilford Press.

Bostrom, R., Waldhart, E. (1978). *The Kentucky comprehensive listening skills test.* Lexington, KY: Kentucky Listening Research Center.

Brown, J., Palmer, A. (1988). *The listening approach: methods and materials for applying Krashen's input hypothesis.* London: Longman.

Burgoon, J. (1985). Nonverbal Signals. In Knapp, M., Miller, G., (eds.), *Handbook of interpersonal communication*, p. 344-390. Beverly Hills: Sage Publications.

Byrnes, H. (1987). Features of pragmatic and sociolinguistic competence in the oral proficiency interview. In Valdman, A., (ed.), *Proceedings of the symposium on the evaluation of foreign language proficiency*, p. 167-177. Bloomington: Indiana University.

Byrnes, H. (1984). The role of listening comprehension: a theoretical base. *Foreign Language Annals*, (17), 317-329.

Canale, M. (1983). On some dimensions of language proficiency. In Oller, J., *Issues in Language Testing Research*, p. 188-202,. Rowley, MA: Newbury House.

Canale, M., Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, (1), 1-47.

Carroll, J. (1958). A factor analysis of two foreign language aptitude batteries. *Journal of General Psychology*, (59), 3-19.

Carroll, J., (1993). *Human cognitive abilities:a survey of factor analytic studies*. Cambridge: Cambridge University Press.

Carroll, J. (1967). Foreign language proficiency levels attained by language majors near graduation from college. *Foreign Language Annals*, (1), 131-151.

Carroll, J. (1962). Prediction of success in intensive foreign language training. In Glaser, R. (ed.) *Training Research and Education*. Pittsburgh: University of Pittsburgh Press.

Carroll, J., Sapon S. (1959). *Modern Language Aptitude Test*. New York: Psychological Corporation.

Cermak, L. (1972). *Human memory: research and theory*. New York: Ronald Press Company.

Cermak, L., Craik F., (1979). *Levels of processing in human memory*. New York: Lawrence Erlbaum.

Child, J. (1987). Language proficiency levels and the typology of texts. In Byrnes, H., Canale, M., (eds.), *Defining and developing proficiency: guidelines, implementations, and concepts*, p 97-106, Lincolnwood , IL: National Textbook Company.

Chomsky, N. (1959). Review of *Verbal Behavior* by B. F. Skinner. *Language*, 35:26-58.

Child, J., Clifford, R., Lowe, P. (1993). Proficiency and performance in language testing. *Applied Language Learning*, 4:1, 19-54.

Clark, J. (1986). *A study of the comparability of speaking proficiency interview ratings across three government language training agencies*. Washington: Center for Applied Linguistics.

Clark, J., Clifford, R. (1987). The FSI/ILR/ACTFL Proficiency scales and testing techniques: development, current status, and needed research. In Valdman, A., (ed.), *Proceedings of the symposium on the evaluation of foreign language proficiency*, p. 1-18. Bloomington: Indiana University.

Cody, M., McLaughlin, M. (1985). Situation as a construct in interpersonal communication research. In Knapp, M., Miller, G., (eds.), *Handbook of interpersonal communication*. p. 263-312. Beverly Hills: Sage Publications.

Collins, A., Loftus E. (1975). A spreading activation theory of semantic processing. *Psychological Review*, (82), 407-428.

Cottrell, G. (1994). Connectionist approaches to natural language processing. Article in Asher, R., Simpson, J., (eds.), *Encyclopedia of Languages and Linguistics*, Vol. 2, p 698-706. Oxford: Pergamon.

Crookall, D., Saunders D. (1989). *Communication and simulation: from two fields to one theme*. Philadelphia: Multilingual Matters.

Cummins, J. (1983). Language proficiency and academic achievement. In Oller, J., *Issues in Language Testing Research*, p. 188-202. Rowley, MA: Newbury House.

Dandonoli, P., Henning, G. (1990). An investigation of the construct validity of the ACTFL proficiency guidelines and oral interview procedure, *Foreign Language Annals*, (23), 11-22.

de Jong, J., (1994). (untitled guest speaker presentation at Language Aptitude Invitational Symposium.)

Department of Defense (1976). *The English Grammar Recognition Test*. Presidio of Monterey: Defense Language Institute.

Department of Defense (1977). *Defense Language Aptitude Battery*. Presidio of Monterey: Defense Language Institute.

Department of Defense (1985). *Technical supplement to the counselor's manual for ASVAB-14*. North Chicago IL: Military Entrance Processing Command.

Doff, A., Jones, C. (1980). *Feelings: A course in conversational English*. Cambridge: Cambridge University Press.

Dunkel, P. (1991). Listening in the native and second/foreign language: toward an integration of theory and practice. *TESOL Quarterly*, (25)3, 431-457.

Dunkel, P., Henning, G., Chaudron, C. (1993). The assessment of a listening comprehension construct. *Modern Language Journal*, 77(2), 180-191.

Dunkel, P., Pialosi, F. (1982). *Advanced listening comprehension: developing aural and note-taking skills*. Cambridge: Newbury House.

Educational Testing Service (1992). *ETS Test Information Brochure No. 42*, Princeton: NTE Programs: School Guidance and Counseling.

Educational Testing Service (1992). (information pamphlet), *Core Battery Tests*, Princeton: NTE Programs.

Faerch, C., Kasper, G. (1987). *Introspection in second language research*. Cleveden, England: Multilingual Matters.

Fahlman, S. (1981). Representing implicit knowledge. In Anderson, J., Hinton, G., *Parallel models of associative memory*, p. 145-160. Hillsdale: Lawrence Erlbaum.

Feyton, C. (1991). The power of listening ability: an overlooked dimension in language acquisition. *Modern Language Journal*, (75), 173-80.

Findler, N.,(ed.), (1979). *Associative networks: representation and use of knowledge by computers*. New York: Academic Press.

Flanagan, J. (1960). *Flanagan Industrial Tests: Expression*. Chicago: Science Research Associates.

Gardner, R., Lalonde, R., Moorcraft, R., Evers, F. (1985). Second language learning: correlational and experimental considerations. *Language Learning*, 35(2), 207-272.

Gardner, R., Lalonde, R., Pierson, R. (1983). The sociocultural model of second language acquisition: an investigation using LISREL causal modeling. *Journal of Language and Social Psychology*, 2 (1), 1-15.

Glen, E. (1989). A content analysis of fifty definitions of listening. *Journal of the International Listening Association*, (3), 21-31.

Goffman, E. (1983). *Forms of talk*. Philadelphia: University of Pennsylvania Press.

Greeno, J. (1970). How associations are memorized. In Norman, D. , *Models of human memory*, p 257-304. New York: Academic Press.

Grice, H. (1975). Logic and conversation. In Cole, P., Morgan, J. (eds.), *Syntax and semantics: speech acts, Vol. 3*, p. 41-58. New York: Academic Press.

Galvin, K. (1985). *Listening by doing--developing effective listening skills*. New York: National Textbook Co.

Hadley, A. (1993). Proficiency-oriented learning: origins, perspectives, and prospects. In Phillips, J.,Diaz, J., *Reflecting on proficiency from a classroom perspective*. Lincolnwood IL: National Textbook Company.

Hatch, E., Yoshitomi, A. (1993). Cognitive processes in language learning. In Hadley, A., *Research in Language Learning: Principles, Problems, Prospects*. Lincolnwood IL: National Textbook Company.

Hatch, E., Shirai, Y., Fantuzzi, C. (1990). The need for an integrated theory: connecting modules. *TESOL Quarterly*, 24 (4), 697-716.

Haycraft, B., Lee, W. (1982). *It depends on how you say it: dialogues in everyday social English*. Oxford: Pergamon:

Heileman, L., Kaplan, I. (1985). Proficiency in practice: the foreign language curriculum. In James, C., (ed.), *Foreign Language Proficiency in the Classroom and Beyond*, p. 55-78. Lincolnwood, IL: National Textbook Company.

Henning, G. (1987). *A guide to language testing*. Cambridge MA: Newbury House.

Higgs T., Clifford, R. (1982). The push toward communication. In Higgs, T., (Ed.), *Curriculum competence and the foreign language teacher*, p 57-80. Skokie IL: National Textbook Company.

Hintzman, D. (1978). *The psychology of learning and memory*. San Francisco: W.H. Freeman and Co.

Horn, J., Stankov, L. (1982). Auditory and visual factors of intelligence. *Intelligence*, (6), 165-185.

Hull, C. (1943). *Principles of Behavior*. New York: Appleton.

Hymes, D. (1972). On communicative competence. In Pride, J., Holmes, J. (ed.), *Sociolinguistics*, p. 269-293. London: Penguin Books.

Jackson, G., (ed.), (1994). *Final report series for the Language Skill Change Project (Staffing Edition)*. Presidio of Monterey: Defense Language Institute Foreign Language Center.

Jacobs, S. (1985). Language. In Knapp, M., Miller, G., (eds.), *Handbook of interpersonal communication*. p. 313-343. Beverly Hills: Sage Publications.

James, C., (ed.), (1985). *Foreign language proficiency in the classroom and beyond*. Lincolnwood, IL: National Textbook Company.

Janssen, C., Hansen, C., Buck, G., DesBrisay, M., Fox, J., Shohamy, E. (1993). *Writing your own language tests*. (Workshop presented at 1993 TESOL convention.)

Karlin, J. (1942). A factorial study of auditory function. *Psychometrika*, (7), 251-278.

Kolodner, J. (1984). *Retrieval and Organizational Strategies in Conceptual Memory: A Computer Model*. Hillsdale: Lawrence Erlbaum.

Knapp, M., Miller, G., (eds.), (1985). *Handbook of interpersonal communication*. Beverly Hills: Sage Publications.

Kramsch, C. (1987). Response to Heidi Byrnes. In Valdman, A., (ed.), *Proceedings of the Symposium on the Evaluation of Foreign Language Proficiency*, p. 183-187. Bloomington: Indiana University.

Kass, R., Mitchell K., Grafton F., Wing, H. (1983). Factorial validity of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 8, 9, and 10. *Educational and Psychological Measurement*, (43), 1077-1087.

Krashen, S. (1987). *Principles and practice in second language acquisition*. London: Prentice Hall.

Lett, J., Thain J. (1994). *The Defense Language Aptitude Battery: what is it and how well does it work?* Paper delivered at Language Aptitude Invitational Symposium.

Long, D. (1989). Second language listening comprehension: a schema-theoretic perspective. *Modern Language Journal*, (73), 32-40.

Lowe, P. (1982). *ILR Handbook on Language Proficiency Testing*. Presidio of Monterey: Defense Language Institute.

Lowe, P. (1985). The ILR scale as a synthesizing research principle: the view from the mountain. In James, C., (ed.), *Foreign Language Proficiency in the Classroom and Beyond*, p 9-53. Lincolnwood, IL: National Textbook Company.

Lund, R. (1991). A comparison of second language listening and reading comprehension. *Modern Language Journal*, (75), 196-204.

Lund, R. (1990). A taxonomy for teaching second language listening. *Foreign Language Annals*, (23), 105-15.

Magnan, S. (1987). Legal caveats on communicative proficiency testing for graduation requirements or teacher certification. In Valdman, A., (ed.) (1987), *Proceedings of the Symposium on the Evaluation of Foreign Language Proficiency*., p. 45-48. Bloomington: Indiana University.

MacWhinney, B. (1994). *Psycholinguistic issues in the assessment of the subcomponents of language abilities*. Paper delivered at Language Aptitude Invitational Symposium.

MacWhinney, B. Bates E., (eds.), (1989). *The crosslinguistic study of sentence processing*. Cambridge: Cambridge University Press.

McLaughlin, M. (1984). *Conversation: how talk is organized*. Beverly Hills: Sage Publications.

Montague, W. (1977). Elaborative strategies for verbal learning and memory. In Bower, G. *Human memory: basic processes*, p 359-436. San Francisco: Academic Press.

Myers, J., McCauley, M. (1985). *Manual: a guide to the development and use of the Myers-Briggs Type Indicator*. Palo Alto: Consulting Psychologists Press.

Norman, D. (1970). *Models of human memory*. New York: Academic Press.

Norman, D., Rumelhart, D., (1970). A system for perception and memory. In Norman, D. (1970), *Models of Human Memory*, p. 19-64. New York: Academic Press.

Omaggio, A. (1986). *Teaching language in context*. Boston: Heinle and Heinle.

O' Malley, J., Chamot, A. (1989). *Learning strategies in second language acquisition*. Cambridge: Cambridge University Press.

O'Mara, F. (1994). Training approaches for reducing student attrition from foreign language training, in Jackson, G., *Final report series for the Language Skill Change Project (Staffing Edition)*. Presidio of Monterey: Defense Language Institute Foreign Language Center.

O'Mara, F., Thain, J. (1994). *Improving the measurement of language aptitude: a psychometric analysis of the Defense Language Aptitude Battery*, paper delivered at Language Aptitude Invitational Symposium.

Oxford, R. (1990). *Language learning strategies: what every teacher should know*. New York: Newbury House.

Petersen, C., Al-Haik, A. (1976). The development of the Defense Language Aptitude Battery (DLAB). *Educational and Psychological Measurement*, (36), 369-380.

Pimsleur, P. (1966), *Pimsleur Language Aptitude Battery*. Psychological Corporation: San Antonio.

Richards, J. (1983). Listening comprehension: approach, design, and procedure. *TESOL Quarterly*, (17), 219-240.

Rost, M. (1990). *Listening in Language Learning*. New York: Longman.

Rutherford, W., Smith M. (eds.), (1988). *Grammar and second language teaching: a book of readings*. Boston: Heinle and Heinle.

Silva, J., White L. (1993). Relationship of cognitive aptitudes to success in foreign language training, *Military Psychology*, 5:2, 79-93.

Skinner, B. (1957). *Verbal Behavior*. New York: Appleton.

Spearitt, D. (1962). *Listening comprehension--a factorial analysis*. (ACER Research Series No. 76). Melbourne: Australian Council for Educational Research.

Stankov, L., Horn, J. (1980). Human abilities revealed through auditory tests. *Journal of Educational Psychology*, (72), 21-44.

Swaffar, J., Bacon S. (1993). Reading and listening comprehension: perspectives on research and implications for practice. In Hadley, A., *Research in Language Learning: Principles, Problems, Prospects*. Lincolnwood IL: National Textbook Company.

Tannen, D. (1982). Oral and literate strategies in spoken and written narratives, *Language*, (58), 1-21.

Tannen, D. (1982). The oral-literate continuum in discourse. In Tannen, D. (ed.), *Spoken and written language: exploring orality and literacy*, p 1-16. Norwood, NJ: Ablex.

Thain J., Lett, J. (1991). *The Learning Strategies Project status report and CY92 Plan* (ESR Report 91-02). Presidio of Monterey: Defense Language Institute Foreign Language Center.

Upshur, J, Homburg, T. (1983). Some relations among language tests at successive ability levels. In Oller, J., *Issues in Language Testing Research*, p. 188-202,. Rowley, MA: Newbury House.

Ur, P. (1984). *Teaching listening comprehension*. Cambridge: Cambridge University Press.

Valdman, A., (ed.) (1987). *Proceedings of the Symposium on the Evaluation of Foreign Language Proficiency*. Bloomington: Indiana University.

Warren W., Nicholas, D. Trabasso, T. (1979). Event chains and inferences in understanding narratives. In Freedle, R., (Ed.), *New directions in discourse comprehension*. Norwood, NJ: Ablex.

Watson, J. (1924). *Behaviorism*, Chicago: University of Chicago Press.

Watson, K., Barker L. (1987). *Watson-Barker Listening Test Video Version*. Auburn, AL: Spectrum.

Wenden, A., Rubin J. (eds.) (1985). *Learner strategies in language learning*. Englewood Cliffs: Prentice-Hall.

Wolvin A., Coakley, C. (1992). *Listening, (4th edition)*. Dubuque IA: Wm. C. Brown.

PSYCHOLINGUISTIC ISSUES IN THE ASSESSMENT OF THE SUB-COMPONENTS OF LANGUAGE ABILITIES[1]

Brian MacWhinney
Carnegie Mellon University

What are the roots of foreign language learning skills? Why are some students successful language learners and others not? Are some languages easier for a certain type of student and other languages easier for another? Answers to these questions could help us make better predictions regarding the outcomes of foreign–language instruction for different combinations of students and languages.

Let us consider a concrete example of how an understanding of individual differences can be used in a particular practical context—that of testing in the context of the instruction conducted at the Defense Language Institute Foreign Language Center (DLIFLC). The primary evaluation instrument used at the DLIFLC is the Defense Language Aptitude Battery or DLAB. The DLAB is a state–of–the–art language test used both for selection of individuals for foreign–language training and for assignment of students to languages. The various languages taught at DLI can be grouped into four categories, ranging in difficulty from easiest to most difficult. DLAB scores are used to ensure that only students with the highest measured language ability will be assigned to the most difficult languages. Assigning weaker students to the harder languages would be a mistake, since the dropout rate would become intolerably high.

This use of the DLAB treats language–learning ability as a unidimensional variable, which we can refer to as 'L'. The more 'L' as student has, the more confident we are that the student can succeed with even a difficult language. However, it makes more sense from a psycholinguistic viewpoint to think of the learner as having a range of abilities—L1, L2, L3, etc.—which share some common variance, but which are also partially dissociable. The other crucial variable determining the success of foreign–language learning is the relative complexity or difficulty of the language being learned. We can combine the psycholinguistic study of individual differences in language learning abilities with contrastive linguistic analyses to build a theory of skill–language interactions that would serve as the basis for successful language–specific prediction of the outcome of foreign language instruction. In building this theory, we need three things:

1. **Skill Analysis.** We would need to have good measures of the learner's strengths and weaknesses across a wide array of tasks. These measures should be based on a thorough psycholinguistic analysis of the basic cognitive, motivational, perceptual, and linguistic skills used in language learning.
2. **Task Descriptions.** We need good contrastive linguistic descriptions that outline the type of demands that particular phonological, morphosyntactic, and discourse structures can place on language learning skills.
3. **A Framework.** Finally, we need a theoretical framework that can allow us to predict and understand interactions between individual skills and target language structures.

---

[1] This paper was prepared with the support of a PRC contract with the Defense Language Institute Foreign Language Center (DLIFLC). The framework for the questions posed here was provided by earlier work assessing the DLAB conducted by Frank O'Mara of PRC and John Thain and John Lett of the DLIFLC. My thanks to each of them for help in understanding the overall context of language testing and the details of the data yielded by the ongoing psychometric analyses of the DLAB.

This type of information can be used to improve (1) the selection of students as candidates for the DLIFLC, and (2) assignment of students to languages.

The initial framework for understanding learner–language relations can be provided by the standard ANOVA model with its main effects and interactions. We begin by recognizing the fact that much of the variance in the outcome of language learning is the results of main effects for the learner and the language. One main effect is based on the overall language–learning abilities of the learner. For any two learners (Le1 and Le2), there can be a main effect for the difference:

$$Le\ 1 > Le2$$

This is to say that if a particular learner Le1 is generally better at learning languages than another learner Le2, we would expect Le1 to surpass Le2 across a wide variety of language–learning experiences. Similarly, for any two languages (La1 and La2) there can be a main effect for the difference:

$$La1 > La2$$

In one such possible ordering, the easiest languages are the Group I languages of Western Europe which use the Roman alphabet, share many cognates with English, and use Indo–European grammatical categories and structures not too terribly different from those of English. Like English, most Group I languages have greatly simplified the original complex grammatical system of Indo–European. In Group II, we find more challenging Indo–European languages. These languages also use the Roman alphabet, but preserve much of the complex grammar of Indo–European. Languages like Lithuanian, German, Romanian, and Hindi are languages of this type. In Group III, at the next level of difficulty, we find those Indo–European languages which maintain both a complex grammar and a non–Roman writing system. These include Greek, Russian, Serbian, and Persian. In Group III, we can also include those 'easy' non–Indo–European languages, such as Hungarian, Tagalog, and Turkish that use Roman characters. In this group, we also find some of the isolating languages of Southeast Asia such as Thai or Vietnamese. In Group IV, we can place non–Indo–European languages with non–Roman orthographies and complex grammatical systems, including Arabic, Japanese, and Korean. Finally, in Group V, we find even more exotic languages like Eskimo (Fortescue, 1984), Warlpiri (Bavin, 1992), Navajo, or Georgian (Imedadze and Tuite, 1992) which present the learner with major challenges in lexicon, grammar, and underlying conceptual organization.

If all prediction of the outcomes of language instruction were the result of these main effects, we would expect to see patterns of this type:

$$Le1La1 > Le1La2 \text{ and } Le2L1 > Le2La2 \text{ (because } La1 > La2)$$
$$Le1La1 > Le2La1 \text{ and } Le1La2 > Le2La2 \text{ (because } Le1 > Le2)$$

However, if there are interactions between learners and languages, we would expect to see two types of reversal patterns:

1. Learner reversals which take the form: $Le1La1 > Le2La1$ but $Le1La2 < Le2La2$

Here the normal learner order is retained for La1, but reversed for La2 because Le1 has particular problems learning this type of language.

2. **Language reversals** which take the form: Le1La1 > Le1La2 but Le2La1 < Le2La2

Here the normal language order is retained for Le1, but reversed for learner Le2 who seems to do particularly well in learning the more difficult language La2.

In practice, language reversals probably tend to occur only between languages that are closely matched in difficulty level and only for learners who are also close in ability levels. For example, we might find that a somewhat stronger learner does better than a slightly weaker learner on Spanish, but not on French, where the somewhat weaker student has some unique affinity for the French sound system. However, we would be extremely surprised to find that a language learner who did an excellent job learning Korean but had no luck at all in learning Spanish.

Although marked language reversals or learner reversals may be rare, reversals in terms of finer levels of detail may be more common. These two additional types of reversals include:

3. **Stage reversals** which take the form: Le1 St1 > Le2St 1 but Le1 St2 < Le2St2

In this type of reversal, Le1 is generally better than Le2. This is true at stage 1 (St1) of language-learning, but at some later stage (St2), Le2 suddenly shows a learning advantage, at least for the material being learned at that stage.

4. **Skill reversals** which that the form: Le1Sk1 > Le2Sk2 but Le1Sk2 < Le2Sk2

In this type of reversal, Le1 is generally better than Le2. However, this is not true across all skills, since the advantage is reversed for skill 2 (Sk2).

It is likely that skill reversals are the underlying causes of stage reversals. For example, it could be that a learner who is good at picking up vocabulary items will do well at the beginning of language learning when vocabulary is so important, but less well at later stages of learning.

A standardized test like the DLAB will do a good job of picking up basic rank orderings among learners, and a thorough contrastive linguistic analysis of a group of languages can be used to establish main effects for languages. However, if we want to improve our ability to predict stage reversals and skill reversals, we will need more fine-grained psycholinguistic measures of language-learning skills as they are applied during the various stages of language instruction.

In the sections that follow, I will suggest some areas that need to be explored in order to better predict each of these four types of reversals.

Individual Differences in Language Processing

Language is the most complex of all human behaviors. At any given moment during language processing, we may be engaged simultaneously in speaking, hearing, reading, formulation, and comprehension. Each of these individual component skills requires the involvement of large areas

of the brain and a complex interplay of local neural processing, functional neural circuits, and high–level strategic organization. Work in cognitive neuropsychology has allowed us to identify some of the basic functions of brain areas in terms of language processing. The use of new scanning techniques in studies of individuals with brain lesions and other language impairments is helping us to understand some of these interactions in terms of the functions of local areas and ways in which local areas are linked together into functional neural circuits for language processing.

Local Processing

In terms of basic–level processing, we know that the temporal lobe has primary responsibility for auditory processing, that the motor strip at the posterior margin of the frontal lobe controls articulation, and that somatosensory input is processed through the sensory strip opposite the motor strip (Goodglass and Geschwind, 1976; Damasio and Damasio, 1988). The cerebellum compiles articulatory gestures from the motor cortex into specific muscle commands. There is a wide area of cortex around the Sylvian fissure and in posterior segments of the frontal lobes where damage can lead to language impairments (Damasio, 1981). Research has pointed toward marked individual differences in such basic attributes as the speed of neural transmission, activation of neural transmitters, involvement of the thalamus and hippocampus in memory and attention, and patterns of neural connectivity.

Commitment and Plasticity

Studies of the development of lateralization during childhood (Farmer, et al., 1991; Aram and Eisele, 1992) indicate that brain areas become progressively committed to particular functions over the course of development. Early in development, the child may lose large areas of cortex, or even an entire hemisphere, and language will still develop normally. As basic linguistic functions develop, they become confined to a smaller area of neural tissue. This leads to an increase in automaticity and speed of processing, but a decline in plasticity and some loss in the potential to function after brain injury. There is also reason to believe that the process that leads to a separation between different languages in bilinguals and second–language learners may also require a commitment of specific neural areas. The plasticity required for these various types of reorganization declines progressively through childhood and adolescence and may be the primary cause of some of the difficulties that adults face in second–language learning.

Integrative Circuits

Current models of the consolidation of episodic memories (Squire, 1992) focus on the role played by the hippocampus (Schmajuk and DiCarlo, 1992; Squire, 1992) in forming higher–level bindings between local areas. In terms of language learning, these bindings allow a variety of local areas to form a series of impressions of the various sensory and conceptual aspects of an utterance or phrase which are then linked together into a new grammatical form or construction. The connections between the hippocampus and local areas are ones used in all mammalian species. However, their use to support language learning is unique in humans and may be supported by other mechanisms. In addition to the hippocampal memory consolidation circuit, there are probably a variety of fairly local circuits that are used in analyzing and breaking apart local memories through a process called 'masking' that has been studied by Cohen and Grossberg (1987). Masking circuits involve the copying of linguistic forms that have been detected

successfully to temporary local buffers so that the system can focus its attention on new incoming material that has not yet been fully processed, while still retaining the recognized material in local memory.

Functional Neural Circuits

The types of local integration supported by the hippocampal episodic system and the local masking system are complemented by a variety of other 'functional neural circuits' that integrate across wider areas of the brain. A prime example of such a circuit is the phonological rehearsal loop (Gupta and MacWhinney, 1995) which links together the auditory processing in the superior marginal gyrus of the temporal lobe with attentional and motor processing from dorsolateral prefrontal cortex. We use this loop to store and repeat a series of words or to speed the learning of new words. There is good reason to believe that the rehearsal loop plays a central role in both first and second–language learning. Moreover, we can use the immediate serial recall (ISR) test to estimate the short–term memory (STM) capacity of this rehearsal loop. Differences in the abilities of learners to store items in this loop have been shown to correlate well with differential success in both first (Gathercole and Baddeley, 1989b; 1990; Gathercole, Adams, and Hitch, 1994) and second (Harrington, 1992) language learning. Other functional neural circuits are involved in basic linguistic activities such as imitation, shadowing, simultaneous translation, speech monitoring, and utterance formulation.

Strategic Control

Finally, it is important not to underestimate the extent to which brain functioning is modified, amplified, integrated, and controlled by higher–level strategic processes. These higher–level processes include mood control, attentional control, motivational control, learning to learn, representational remapping, promotion of analogies, and applications of scripts (Naiman, et al., 1978; O'Malley and Chamot, 1990). The degree to which the foreign language learner can use phonemic recoding (Perfetti, Bell, and Delaney, 1988), graphemic visualization, translational equivalents (MacWhinney, 1992), and vocal tract models to facilitate language learning will determine relative success or failure across a wide range of foreign–language skills at various stages in language learning.

Level of Attention

Some learners pay more attention to overall conceptual structure, attempting to process sentences through top–down inferential processes (Bransford and Franks, 1971; Bransford, Barclay, and Franks, 1972; Barclay, et al., 1974; Kintsch, 1977; VanDijk and Kintsch, 1983; and Lombardi and Potter, 1992), whereas other learners focus more on listening to phonetic detail (Flege, Takagi, and Mann, 1995). It is easy to believe that those learners who pay more attention to phonetic detail in listening will acquire better phonological control over the language, but there is no research directly supporting this intuition.

Monitoring

One learner variable that has been shown most clearly to correlate with higher achievement is the use of monitoring or error–checking. Students who attempt to detect and correct their own errors and who make productive use of feedback from their instructors tend to perform better on achievement tests (Carroll and Swain, 1992). Such findings seem to contradict claims about the

importance of disengaging the Language Monitor (Krashen, 1978; 1982) as well as claims about the marginal role of negative evidence (Pinker, 1989). It is also true that learners appear to be differentially sensitive to instruction in the use of good language-learning strategies. If a learner is open to instruction in the use of these strategies, it is not important that the strategies be fully learned and controlled before the beginning of second-language instruction. However, the measurement of the learner's openness to such instruction could be an extremely difficult psychometric problem.

How might these various factors impact the learning of languages differentially? Could it be that learning of 'difficult' languages such as Chinese and Arabic, requires learners who make maximum use of learning strategies, whereas learning of 'simple' languages such as Dutch and Spanish requires little use of language-learning strategies? Although such a relation seems plausible, we do not yet have any data that could allow us to evaluate such hypotheses.

Native-Language Skills

Psychometrically speaking, the simplest model of individual differences in language learning would predict success in second-language learning entirely on the basis of skills that had already been demonstrated in native-language learning. A learner who is fast at processing words in the native language should also be fast at processing newly learned words in the second language. A learner who is good at comprehending complex passages in the native language should eventually be good at comprehending complex passages in the second language. However, there are a variety of reasons to expect that direct prediction of second-language individual differences from native-language individual differences will be far from absolute. Tests of native-language abilities tend to measure the results of the application of these skills, rather than the skills themselves. In the years intervening between basic native-language acquisition and the beginning of second-language learning, these skills may have fallen into disuse or may have atrophied altogether (Werker, et al., 1981; Johnson and Newport, 1989; 1991;). In fact, as basic native-language skills become solidified through neural commitment of local areas, the brain's capacity to add new material to the processing in these areas diminishes. However, if we turn to those higher-level integrative processes that are not supported by specific local areas, the prediction of second-language learning on the basis of native-language skill may be more successful. For example, we might expect that a learner who has a rich ability to process strings of words in the articulatory loop will be able to use this ability to support foreign-language learning.

At still higher levels of language learning, prediction of second-language attainment on the basis of native-language abilities should be fairly powerful. For example, we might well expect that a person who is a successful public speaker in the native language would also be a successful public speaker in the foreign language. We could measure a learner's control of narrative, argumentation, poetry, genre variations, literary criticism, and scientific writing in the native language as an excellent way of predicting eventual control of similar structures in the second language. At the same time, it is likely that the nature of the learner's overall attitude toward the native language will have a great influence on second-language learning. There is a wide variety of behaviors that can reflect a fascination with language use and language learning. These include interest in dictionaries, crossword puzzles, conversation, novels, plays, debates, stories, jokes, and all other forms of verbal entertainment and analysis. Positive experience with these forms in the

native language can be generalized to early language-learning experiences in the second language. In addition, the learner may realize that successful learning occurs best when these positive experiences are maximized. Having learned one foreign language successfully, these same strategies can then be reapplied in increasingly successful ways.

Autosupport

Together, we can think of these various high-level strategies as forming a system for 'autosupport' that is crucial to adult second-language learning. The young child benefits directly from two fundamental supports for language learning. The first is the presence of a fresh, uncommitted neurological basis. The second is the provision by the child's caretakers of a rich system of social support. Parents read story-books to children, ask questions and wait patiently for answers, and provide names for unfamiliar pictures. No adult receives this immensely supportive scaffolding for language learning. Instead, the adult learner must compensate for the loss of these support systems by generating 'autosupport' mechanisms that rely on more complex functional neural circuits. Despite the massive individual differences that evidence themselves in native-language learning, nearly all children learn language. The same is not true for second-language learning, where the absence of good acquisition of autosupport strategies can lead to total failure in second-language learning, even when the practical negative consequences of this failure are enormous.

The applications of autosupport strategies allow the adult language learner to compensate for two types of handicaps. On the one hand, the adult learner must work to gain the richness of exposure to primary language data that the young child gets for free. On the other hand, the adult must fight an uphill battle against the commitment that has occurred in local areas of neural processing. The young learner has access to large amounts of uncommitted and fresh neural tissue, whereas the older learner works against direct competition from older, well established structures. The great wonder of adult second-language learning is the fact that learning can occur at all. The fact that it can testifies to the importance of input maximization, the residual capacity of the brain, and the ways in which functional neural circuits can be used to retune local processing areas.

Testing

Psycholinguistic research has devoted a great amount of attention to the testing and measurement of these underlying skills and processes. Measures of articulatory control (MacNeilage, 1970), auditory sensitivity (Tallal and Stark, 1980), baseline reaction speed (Kail, 1992), decision speed, choice speed, short-term memory capacity, rehearsal capacity, sentence span (Daneman and Carpenter, 1980), analogistic processing (Gentner, 1988), retrieval speed (Kilborn, 1989), retrieval accuracy, motivational factors, and attitudes toward language have all received extensive attention in the psycholinguistic literature. However, few of these measures exist in forms that can applied in the context of paper-and-pencil tests, since many of them look at online processing in the context of reaction-time studies.

Orthographic Learning

Having surveyed some of the basic mechanisms supporting language learning, we next consider ways in which linguistic structures can emerge as major roadblocks to progress during language learning. One such area of potential roadblocks is the learning of new and difficult orthographic systems.

There are two major areas of orthographic difficulty that can confront a foreign language learner. The first dimension is the presence of irregularities and inconsistencies in phoneme–grapheme correspondences. The second dimension is the presence of a new set of orthographic characters in the foreign language. For the English–speaking learner, this means the use of non–Roman orthographies, as well as special diacritic markings. Moreover, these two factors can also interact, since non–Roman orthographies can also be irregular in their mappings of phonemes to graphemes. A language like Spanish poses virtually no major orthographic difficulties to an English–speaking learner, whereas a language like Chinese presents the learner with an enormous orthographic learning task.

Phoneme–Grapheme Regularities

Learning to read and spell words in a new language can involve learning of a complex set of spelling patterns and rules. Languages like Polish, Hungarian, and Spanish are extremely consistent in their use of particular letters to mark particular phonemes. These languages tend to use a single letter to mark a single phoneme, leading to a consistent mapping from phonemes to graphemes, as well as from graphemes to phonemes. Languages like German or Dutch show consistency in the mapping of clusters of graphemes to phonemes, but a fair amount of indeterminacy in the mapping of phonemes to graphemes. In these languages, you know how to pronounce a new word if you see it spelled, but you are not sure how to spell a new word if you hear it pronounced. Other, even more difficult, languages, like English and French, tolerate a huge amount of plurifunctional marking and irregular patterns. The factor that is involved in these variations is the regularity of the phoneme–grapheme correspondences in the language (Venezky, 1970).

Simplicity of Mapping

When we move outside the realm of Roman–based orthographies, we find a wide variety in the shapes of the orthographic systems confronting the learner. The basic psycholinguistic principal operative is one of preference for one–to–one mappings. Ideally, the learner wants to find one non–Roman character for each character of the Roman alphabet. To the extent that this can be done, learning is facilitated. In order to read a new word, one takes a character in the new language, translates it to a character in English Roman script and then activates the corresponding phoneme. Eventually this mediation through Roman characters and English phonemes is dropped and the mapping from graphemes to phonemes is reconstructed in the new language. However, it would be a mistake for teachers to think that the mapping can be learned directly right from the beginning. Rather, it is likely that learners who can move quickly through the period of transfer and remapping from a Roman base will be those who are quickest to master the new orthography. Similarity of mapping. In Greek and Cyrillic, the mapping of characters to the Roman system is fairly transparent. Some of the letters even share a few physical characteristics. These iconic

relations provide initial retrieval cues to the learner during the acquisition of the new alphabet. However, orthographies such as those of Hebrew, Indian *devanagari*, or Arabic, have no clear mapping to Roman characters. A comparison of the learning of scripts like *devanagari* with the learning of Cyrillic would help to illuminate the actual importance of script similarity within the context of different Indo–European languages. For a procedure that can be used to illuminate these functions crosslinguistically, see Kempe and MacWhinney (1994).

Nonphonemic Scripts

Although most orthographies are based on phoneme–grapheme correspondences, systems such as Chinese and ancient Egyptian use characters that have no match to individual sounds. The learning of non–phonemic scripts is impacted by a rather different set of learner variables. The kinds of learner variables we expect to be important here are similar to those that are important in first language word learning and perception. Learners relying on holistic learning are unable to piece together words from phoneme correspondences and must acquire words as phonological wholes. Such learners would do better with systems oriented towards whole words, such as Chinese. Full literate command of a language with difficult spelling patterns can be a tough matter and can set an upper limit on the achievement of a student. Limits of this type are also found in native–language acquisition for some of the rarer *kanji* forms in Japanese. A native–speaker learner may acquire certain *kanji* in high school which he or she then seldom uses again in later life. However, the majority of the world's orthographic systems are analytic or alphabetic in nature, and most learners will need to apply analytic abilities of the type they initially used as children learning the English alphabet and its use in early reading.

Psycholinguistic Considerations

Patterns of phoneme–grapheme irregularities provide us with a good illustration of ways in which learner characteristics can interact with language features. Highly analytic learners should do better with regular languages and less analytic learners should do comparatively better with languages that have irregular systems. Baron and his colleagues (Baron, 1977a; 1977b; 1979; 1980; Baron and Strawson, 1976) have used psycholinguistic methods to classify readers as either 'Phoenician' or 'Chinese' depending on their relative use of analytic versus holistic approaches to lexical and orthographic learning. In terms of this dimension, we would expect analytic learners to do well with regular systems and holistic learners to do comparatively better with more irregular systems and non–phonemic systems.

Psychologists have created a variety of detailed computational models of orthographic processes in reading and spelling. These models have been tested as accounts of deep dyslexia in adults (Coltheart, Patterson, and Marshall, 1987; Plaut and Shallice, 1991; Plaut and McClelland, 1993), lexical decision processes in normal subjects (Kawamoto and Zemblidge, 1992; Kawamoto, 1993), and word learning in children (Seidenberg and McClelland, 1989). Despite disagreements about general approaches, all models in this area must deal with the distinction between learners who emphasize rules and learners who emphasize rote.

Testing

One way of measuring students' abilities to acquire new orthographies would be to simply present a new alphabet that maps English to new characters. The learner would be required to study the alphabet quickly and then use it to identify possible spellings of English words. The alphabet could be either similar to English or radically different with shapes like that of *devanagari* or *hangul*. In addition there could be a set of whole word forms that the student would need to memorize. These would parallel characters in the Chinese system.

Phonological and Phonetic Learning

It is difficult to overestimate the importance of phonological factors in foreign–language learning. Typically, second–language learners who have not received careful phonetic training find it difficult to lose all traces of their native accent, if they have begun acquisition of the foreign language after age 20 (Oyama, 1976; Johnson and Newport, 1989; 1991).

Receptive Phonology

There is evidence that receptive phonological abilities become locked in on the native language even during infancy (Werker, et al., 1981). Lively, Pisoni, and Logan (1990) have shown that even the most difficult phonological contrasts can be learned during adulthood given sufficient practice, but whether it is possible to reach native–level performance across the board is difficult to demonstrate.

Motor Production

On the articulatory side, Hanson–Bhatt has conducted detailed studies of phonological transfer in second–language learners that have demonstrated feature–by–feature transfer from the native language to the second language. These recent analyses serve to update earlier ideas regarding phonological learning developed within the context of contrastive analysis (Lado, 1971). Careful attention to phonetic detail can help learners overcome some of these limitations (Flege, et al., 1995). However, there is little work that would provide clear guidance regarding the nature of those individual differences that contribute to successful acquisition of second–language accent. In particular, we do not know whether phonological acquisition of a new language is impeded primarily by ossification of the perceptual system or primarily by difficulties in establishing new procedures for controlling motor output (MacNeilage, 1970; Liljencrants and Lindblom, 1972; McNeil and Kent, 1990; Odell, et al., 1991).

Testing

Match–to–sample and same–different tests for prosodic and segmental contrasts are easy ways of testing for ability to perceive phonological contrasts. On the articulatory side, measures used in the field of speech and language disorders, such as rapid syllable repetition rate or syllable shadowing, could be adapted for use in the foreign language–learning context.

Lexical Learning

Perhaps the single biggest task facing the language learner is the acquisition of new words. In order to develop even moderate fluency in a new language, the learner must acquire several thousand new lexical items. Lexical learning involves three basic processes: form learning, function learning, and the establishment of retrieval cues that promote the association of form to function (Keenan and MacWhinney, 1987).

The Phonological Loop

The work of Baddeley and associates (Baddeley, 1986; 1992; Baddeley, Papagno, and Vallar, 1988; Gathercole and Baddeley, 1989a; 1990; Papagno, Valentine, and Baddeley, 1991; and Gathercole, et al., 1992) has underscored the role of articulatory rehearsal in word learning. There is evidence that this loop is used during word learning as well as during immediate serial recall (ISR). Gupta and MacWhinney (1994; 1995) have argued that the loop is based on a specific neural circuit connecting lexical phonological representations in posterior cortex and output forms in anterior cortex. Children with specific language impairment (SLI) seem to have a deficit in the use of this rehearsal loop (Gathercole and Baddeley, 1989b; 1990). In second–language learning, there is good reason to believe that successful language learners would be those who have a maximally well–developed ability to continue verbal rehearsal.

Phonological Processes in Rehearsal

It is likely that the process of verbal rehearsal interacts significantly with the shape of phonological coding. In languages with phonological systems that are close to those of English, learners could make productive use of their full rehearsal abilities. However, in languages with more difficult sound systems, there could be a greater load imposed on articulatory rehearsal and therefore a slower rate of word learning (MacWhinney, 1994). In this regard, problems could arise not only from segmental phonology, but also from suprasegmental markings such as vowel and consonant length, as well as tone and stress. As words place a greater and greater load upon the articulatory loop, we will expect to see simplifications and reductions to English–like forms. In this way, phonological difficulties can be reflected in problems of lexical learning.

Semantic Factors

Languages differ even more markedly in the demands they place on semantic aspects of word learning. In the very worst case, learning of a new word cannot depend on anything available from the first language. The new word would involve a complex set of new and difficult phonological mappings and a totally unfamiliar and complex set of semantic meanings. In languages such as Navajo or West Greenlandic, this worst case scenario may often be the actual case. Languages such as Korean or Japanese may be only marginally better. However, in languages closer typologically and culturally to English, there is a variety of factors that can facilitate learning.

There are at least four support factors that can facilitate this learning: cognate mapping, analogic mapping, semantic transparency, and semantic overlap. The best case for the learner is the case of cognate learning. It is obviously much easier to learn Spanish *republica* for 'republic' than Hungarian *nepkoztarsasag*. In cases where there is no direct cognate, there may still be a certain

symmetry between the two languages. Sometimes words with parallel derivational or compound structure across languages are known as 'mirror words'. For example, German *Worterbuch* (words–book) can serve as a reasonable basis for the learning of Hungarian 'szótár' (word–book), whereas English 'dictionary' is much more helpful as a basis for learning Spanish *diccionario*. Even when a word in a new language cannot be perceived as a cognate or a mirror word, it may be relatively easy to decipher its meaning compositionally. For example, it is easy enough to understand that German *zweikeimblattrige* means 'dicotyledon', because the German word can be taken apart as 'two–kernel–leafed'. Or, if the student knows the French word *joie*, it is relatively easy to decipher the meaning of *joyeux* or *joyeusette*.

Semantic Overlap

Even when supports like cognates and semantic transparency are not available, languages may promote lexical learning simply by maximizing the overlap between concepts. For example, the word 'milk' in English means almost exactly the same as the word *Milch* in German. The learner typically begins with the assumption that this overlap is virtually complete. In fact this process is so strong initially, that Kroll and associates (Kroll, 1990; Kroll and Sholl, 1992) have shown that virtually all early lexical learning is mediated through first language concepts. However, for languages such as Korean and Japanese that have words with meanings that are very different from those of English, attempts to transfer meaning can lead to error, and learning itself is often exceedingly incomplete (Ijaz, 1986).

Testing

Given the importance of lexical learning, it is surprising that predictive tests seldom measure of this ability. Kempe and MacWhinney (in press) have developed a test of lexical learning based on the lexical decision task. This test is useful as a measure of early second–language attainment. In order to measure ability quickly to acquire a new set of words, the most obvious test would be one based on the old verbal learning technique of paired–associate learning. In a test of this type, the new words to be learned could be either English–like words or words that resembled those in a new language.

Morphosyntactic Learning

Tests like the DLAB tend to focus on measurement of the skills involved in grammatical learning. These skills certainly constitute an important component of language learning. Let us take a look as some of the component skills involved in grammatical learning and their differential use across languages.

Grammatical Markings

Languages differ markedly in the extent to which they require the learner to pick up large systems of nominal declension and verbal conjugation. At one extreme are languages like Navajo, with rich systems of aspects, person, case, number, and voice—all blended together in intricate phonological alternations in long complex verbs that also mark the shape of the object and properties of the location of the activity and direction of the action in a variety of spatial dimensions. At the other extreme are languages like English, Afrikaans, or Swahili that have only a few affixes and little in the way of obligatory morphological marking of grammatical categories.

The ways in which languages organize their markings of things like tense, number, space, and time (Talmy, 1976; 1977; 1988) are rich and varied (Bloomfield, 1961; Greenberg, 1978). Simplifying enormously, one can reduce this immense complexity to three basic dimensions: marking complexity, class membership complexity, and the complexity of the underlying grammatical categories.

Marking Complexity

In the simplest of grammatical systems, there are very few grammatical markings and the issue of combining of grammatical markings seldom arises. However, even in an analytic language such as English, some combinations can occur. For example, the plural of 'girl' is 'girls' and the possessive of 'girl' is 'girl's'. Combining these two, we might have expected 'girls's', but English prefers brevity and we have only 'girls''. In languages with more category markings, three configurations of categories are available. The most methodical solution is the agglutinative solution which concatenates markers one after another. Good examples of agglutinative languages are Turkish and Quechua. If these markers exercise strong phonological effects on each other, we have polysynthetic systems like Paiute or Greenlandic. The third solution is the fusional solution. In languages such as Latin, a given suffix or article may simultaneously signal three or even four grammatical categories. These distinctions are well–known and there is no need to review them further here. What is more important from the viewpoint of language learning is the distinction between paradigm learning and the learning of formal classes. The evidence currently available indicates that these are separate tasks. For example, in German child language (Mills, 1986) there are very few errors in the learning of case and also few errors in the assignment of nouns to gender class. However, for second–language learners of German, the acquisition of the basic paradigm is very easy but the learning of noun gender is extremely difficult.

Until very recently, the possible existence of individual differences in abilities to learn grammatical systems was totally uncharted territory. Recent work (Gopnik 1990; Gopnick and Crago, 1990; Pinker, 1991; van der Lely 1993; and Van der Lely and Howard,1993) has suggested that some children with language disorders may have a specific disability that blocks them from acquiring grammatical paradigms. Unfortunately, this work is marked by theoretical overstatements and methodological flaws and should not yet be viewed as anything more than suggestive. It may well be the case that some learners have specific problems in the area of inflectional morphology, but the exact nature of these problems remains to be more carefully delineated.

Category Membership

There has been a fair amount of work recently on the learning of grammatical gender in German (MacWhinney, 1978; MacWhinney, et al., 1989; Clahsen and Penke, 1991; Clahsen and Rothweiler, 1992; Clahsen, et al., 1992; and Marcus, et al., 1993). This work has underscored the importance of detailed low–level phonological cues in assigning words to gender class. For example, the ending *–e* is used as a cue to feminine, the ending *–en* as a cue to masculine, and the ending *–chen* as a cue to neuter. Sometimes these cues involve derivational items and sometimes they compete with other cues. There are also important semantic cues such as 'alcoholic beverage', 'stone', or 'superordinate.'

Conceptual Complexity

The formal shapes of paradigms and the membership of specific items in categories can seldom transfer from one language to another during second-language learning. This is certainly true for English learners, whose system of grammatical marking is minimal to begin with. However, the underlying meaning structure of the concepts being expressed by grammar can be transferred from one language to another. Let us compare two different grammatical categories in English and German: plural and dative. The category of plural marking on the noun is quite parallel between the two languages. Neither language has a dual marking. In both there are suffixes to mark plurality. The German system for plural marking is far more complex, but the underlying notion of plurality being expressed is the same as in English. Marking of the German dative, on the other hand, has no real parallel in English. It is true that English uses the preposition 'to' or the double object construction to mark the indirect object. And the student could assume some equivalency between the English indirect object and the German dative. However, this similarity is quite partial. The trade-off between the double object construction and the prepositional dative has no exact match in German. Most importantly, the German dative can also be used to mark the object of certain prepositions and this is in turn conditional upon the nature of the action of the verb. There is also a limited use of the dative in possessives, and there are a number of German forms in which the dative is the experiencer rather than the recipient.

Problems with the conceptual bases of grammatical categories may be some of the crucial determinants of learner problems with 'exotic' languages such as Korean and Japanese. For example, marking of tense or aspect in Japanese or the use of *wa* and *ga* require the learning of new conceptual mappings.

Testing

It is relatively easy to test for learner abilities in the area of paradigm learning and class formation. For example, subsections of the DLAB do a good job measuring these skills. However, it is much more difficult to test for ability to acquire new conceptual structures. One way in which this could be done is through induction of a grammatical category from examples. The contrast in Spanish between *ser* and *estar* could be used as a prototype. It should be possible to present the student with a series of example sentences in which the one form describes a permanent attribute and another form describes a transient quality. If the student can induce new concepts in this context, they would evidence ability to acquire new concepts in the larger language-learning task.

Syntactic Processing and Learning

It is difficult to separate the acquisition of formal marking systems from the overall syntactic system of a language. Perhaps the easiest way to think of the relation is to realize that syntax uses both local morphological markings and non-local word order or configurational patterns to express a variety of underlying concepts and meanings. Chomksy (1981; 1982; 1986) has attempted to characterize syntactic differences between languages in terms of a small set of key parameters, such as treatment of subject pronouns, movement of *wh*-words, and placement of adverbs and other verbal markers. It is difficult to find a single parameter which has received uniform linguistic support. Moreover, the exact role of parameters in language learning is still very unclear (Truscott and Wexler, 1989; Lightfoot, 1989; 1991; Hyams and Wexler, 1993;

Poeppel and Wexler, 1993). Despite these uncertainties, the parameter–setting framework for second–language acquisition syntax has motivated some interesting work, particularly from White and her students (White, 1989; 1990; 1991; 1992; Trahey and White, 1993;). An alternative view of the learning of second–language syntax has been developed within the Competition Model of Bates and MacWhinney (MacWhinney, 1987; MacWhinney and Bates, 1989). The Competition Model emphasizes traditional psychological and psychometric constructs such as transfer, cue strength, cue validity, and processing cost. The Competition Model has been applied to the study of second–language acquisition of grammar in a dozen languages and has made uniformly successful empirical predictions.

A Concrete Example

In order to see how the Competition Model and Chomskyan parameter–setting would deal with a particular aspect of language learning, let us look at the case of the learning of adverb placement. The parameter–setting account of adverb placement grounds learning on the resetting of a parameter for strong AGR marking. This parameter would relate the fact that German places the negative after the verb to its placement of the adverb after the verb. In English, on the other hand, both the adverb and the negative marker precede the verb. English says 'He often watches television' and German says '*Er sieht oft fern*'. White's work with second–language learners shows that instruction focusing on one component of the parameter does not influence learning of the other components. Instead, it appears that each aspect of the syntactic system is learned independently in its own right. This finding matches best with the analysis of the Competition Model. Both the Competition Model and the parameter–setting view assume an initial transfer of word–order patterns from English, and this is certainly what is found. However, parameter–setting requires a linkage between this pattern of learning and other aspects of learning. To date, no strong linkages of this type have yet been empirically confirmed. Given these negative findings and related theoretical problems, it would probably be a mistake at this point to rely on parameter–setting theory as a guide toward elaboration of tests like the DLAB.

Local versus Nonlocal Marking

Studies within the Competition Model framework have suggested another dimension that may be an important determinant of syntactic learning. This is the contrast between local and configurational marking. A clear case of local marking is the use of the Spanish preposition *a* with the direct object. Although this preposition is not formally a case marking, it functions as one in psycholinguistic studies of Spanish sentence processing (Kail, 1989). English learners of Spanish or Italian (Bates and MacWhinney, 1981) may at first attempt to use English word–order strategies to mark the direct object, but they will soon realize that the variable nature of Spanish word order makes this impossible. The prototypical example of a nonlocal marking is the agreement between the verb and the subject. Initially, one might think that languages that use redundant marking of grammatical categories would be somehow easier to learn. However, Competition Model studies of agreement marking in languages such as Hungarian, Arabic, Italian, Spanish, German, Serbo–Croatian, and French have shown that this is not the case. In fact, processing of subject–verb and object–verb agreement cues is one of the most difficult aspects of sentence processing, one which apparently places heavy demands on working memory and phonological rehearsal. Work by Bock and colleagues (Bock and Miller, 1991; Bock and Eberhard, 1993) in English supports this interpretation. Indeed, it appears

that problems in subject–verb agreement marking may be an important dimension to measure as a possible indicator of language-learning limitations. Note, however, that these problems are not so extreme for gender agreement within the noun phrase (Urosevic, et al., 1988), although they do effect gender agreement between the subject and the verb in Arabic, for example.

Testing

Testing could be done using the basic sentence–interpretation task. Test items should be chosen to sample from the various agreement structures and should include both local markings of sentence roles and configuration or word–order markings. A book by MacWhinney and Bates (1989) presents a wide variety of experimental techniques that could be adapted to the study of real–time sentence processing in the second language. Specific studies applying this methodology include McDonald and MacWhinney (1989; 1995). Kilborn (1989) has shown ways in which the imposition of an additional cognitive load through auditory noise or concurrent tasks can reveal deeper processing difficulties in even normal adult native speakers.

Conclusions

This brief survey has examined ways in which language–learning abilities interact with complex linguistic structures. Adult second–language learners face problems using low–level learning mechanisms to acquire the forms of a new language against the interference patterns from the first language. To overcome this, language learners must rely on functional neural circuits, motivational support, and other behaviors under strategic control. It is possible that learners have markedly different profiles of skills and that the interactions of these different profiles with different target languages could produce a variety of stage reversals and skill reversals. In order to understand this possible effect in greater detail, we will need to improve our methods for measuring functional language–learning skills.

It is important to place these potential interaction effects into a broader context. First, we should remember that the largest percentage of the variance in foreign language–learning outcomes will continue to be the main effect based on the overall ability level of the learner and the overall level of difficulty of the language. However, within this general framework, we need to study additional interactions for both practical and theoretical reasons. Secondly, this model of learner–language interactions ignores the other important determinant of the outcome of language learning, which is the nature of the educational treatment. A good teacher may be able to help a good student overcome some particular roadblock during language learning. At the same time, a good learner may be able to make use of the teacher as a resource in the process of overcoming specific disabilities or difficulties.

Aram, D., And Eisele, J. (1992) Plasticity And Recovery Of Higher Cortical Functions Following Early Brain Injury. In F.B.a.J. Grafman (Ed.), *Handbook Of Neuropsychology: Child Neuropsychology*. Amsterdam: Elsevier.

Baddeley, A. (1986). *Working Memory*. Oxford: Oxford University Press.

Baddeley, A. (1992). Working Memory: The Interface Between Memory And Cognition. *Journal Of Cognitive Neuroscience, 4*, pp. 281–288.

Baddeley, A., Papagno, C., And Vallar, G. (1988). When Long–Term Learning Depends On Short–Term Storage. *Journal Of Memory And Language, 27*, pp. 586–595.

Barclay, J.R., Bransford, J.D., Franks, J.J., Mccarrell, N.S., And Nitsch, K. (1974). Comprehension And Semantic Flexibility. *Journal Of Verbal Learning And Verbal Behavior 13*. pp. 471–481.

Baron, J. (1977a). Mechanisms For Pronouncing Printed Words: Use And Acquisition. In D Laberge, And S. Samuels (Eds.), *Basic Processes In Reading: Perception And Comprehension*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Baron, J. (1977b). What We Might Know About Orthographic Rules. In S. Dornic (Ed.), *Attention And Performance VI*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Baron, J. (1979). Orthographic And Word–Specific Mechanisms In Children's Reading Of Words. *Child Development, 50*, pp. 60–72.

Baron, J., And Strawson, C. (1976). Use Of Orthographic And Word–Specific Knowledge In Reading Words Aloud. *Journal Of Experimental Psychology: Human Perception And Performance, 2*, pp. 386–393.

Baron, J., Treiman, R., Wilf, J., And Kellman, P. (1980). Spelling And Reading By Rules. In U. Frith (Ed.), *Cognitive Processes In Spelling* . London: Academic Press.

Baron, R. W. (1980). Visual And Phonological Strategies In Reading And Spelling. In U. Frith (Ed.), *Cognitive Processes In Spelling*. London: Academic Press.

Bates, E., And MacWhinney, B. (1981). Second–Language Acquisition From A Functionalist Perspective: Pragmatic, Semantic And Perceptual Strategies. In H. Winitz (Ed.), *Annals Of The New York Academy Of Sciences Conference On Native And Foreign Language Acquisition*. New York: New York Academy Of Sciences.

Bavin, E. (1992). The Acquisition Of Warlpiri. In D. I. Slobin (Ed.), *The Crosslinguistic Study Of Language Acquisition: Volume 3*. Hillsdale, NJ: Lawrence Erlbaum Associates

Bishop, D. (1981). Plasticity And Specificity Of Language Localization In The Developing Brain. *Developmental Medicine And Child Neurology, 23*, pp. 251–265.

Bloomfield, L. (1961). *Language*. New York: Holt, Rinehart And Winston.

Bock, K., and Miller, C. (1991). Broken Agreement. *Cognitive Psychology, 23*, pp. 45–93.

Bock, K., And Eberhard, K. (1993). Meaning, Sound And Syntax In English Number Agreement. *Language And Cognitive Processes, 8*, pp. 57–99.

Bransford, J., Barclay, R., And Franks, J. (1972). Sentence Memory: A Constructive Vs. Interpretive Approach. *Cognitive Psychology, 3*, pp. 193–209.

Bransford, J. D., And Franks, J. J. (1971). The Abstraction Of Linguistic Ideas. *Cognitive Psychology, 2*, pp. 331–350.

Carroll, S., And Swain, M. (1992). The Role Of Feedback In Adult Second–Language Acquisition: Error Correction And Morphological Generalizations. *Applied Psycholinguistics, 13*, pp. 173–198.

Chomsky, N. (1981). *Lectures On Government And Binding*. Cinnaminson, N.J.: Foris.

Chomsky, N. (1982). *Some Concepts And Consequences Of The Theory Of Government And Binding*. Cambridge, MA: MIT Press.

Chomsky, N. (1986). *Barriers*. Cambridge, MA: MIT Press.

Clahsen, H., And Penke, M. (1991). The Acquisition Of Agreement Morphology And Its Syntactic Consequences: New Evidence On Gertnan Child Language From The Simone–Corpus. *Unpublished Manuscript*. Allgemeine Sprachwissenschaft Universitat Dusseldorf

Clahsen, H., And Rothweiler, M. (1992). Inflectional Rules In Children's Grammars: Evidence From German Participles. In G. Booij, And J. Van Marle (Eds.), *Yearbook Of Morphology*. Dordrecht: Kluwer.

Clahsen, H., Rothweiler, M., Woest, A., And Marcus, G. (1992). Regular And Irregular Inflection In The Acquisition Of German Noun Plurals. *Cognition, 45*, pp. 225–255.

Cohen, M., And Grossberg, S. (1987). Masking Fields: A Massively Parallel Neural Architecture For Learning, Recognizing, And Predicting Multiple Groupings Of Patterned Data. *Applied Optics, 26*, pp. 1866–1891.

Coltheart, M., Patterson, K., And Marshall, J. (1987). *Deep Dyslexia*. London: Routledge.

Damasio, A. R., And Damasio, H. (1988). The Neuroanatomical Correlates Of Aphasia And The Understanding Of The Neural Substrates Of Language. In Hyman, L.M. and a.C.N. Li (Eds.), *Language, Speech And Mind–Studies In Honor Of Victoria A. Fromkin*. Routledge & Kegan Paul.

Damasio, H. (1981). Cerebral Localization Of The Aphasias. In M. T. Sarno (Ed.), *Acquired Aphasia*. New York: Academic Press.

Daneman, M., And Carpenter, P. (1980). Individual Differences In Working Memory And Reading. *Journal Of Verbal Learning And Verbal Behavior, 19*, pp. 450–466.

Farmer, S., Harrison, L., Ingram, D., And Stephens, J. (1991). Plasticity Of Central Motor Pathways In Children With Hemiplegic Cerebral Palsy. *Neurology, 41*, pp. 1505–1510.

Flege, J., Takagi, J., And Mann, V. (1995). Japanese Adults Can Learn To Produce English "r" And "l" Accurately. *Language Learning, 39*, pp. 23-32.

Fortescue, M. (1984). Learning To Speak Greenlandic: A Case Study Of A Two–Year–Old's Morphology In A Polysynthetic Language. *First Language, 5*, pp. 101–114.

Gathercole, S., Adams, A., And Hitch, G. (1994). Do Young Children Rehearse? An Individual–Differences Analysis. *Memory And Cognition, 22*, pp. 201–207.

Gathercole, S., and Baddeley, A. (1989a). Evaluation of The Role of Phonological STM in The Development Of Vocabulary Of Children: A Longitudinal Study. *Journal Of Memory And Language, 28*, pp. 200–213.

Gathercole, S., and Baddeley, A. (1989b). The Role Of Phonological Memory In Normal And Disordered Language Development. In E. I. Lundberg, And G. Lennerstrand (Eds.), *Brain And Reading*. New York: Macmillan.

Gathercole, S., and Baddeley, A. (1990). Phonological Memory Deficits In Language Disordered Children: Is There A Causal Connection? *Journal of Memory and Language, 29*, pp. 33-60.

Gathercole, S., Willis, C., Emslie, H., and Baddeley, A. (1992). Phonological Memory And Vocabulary Development During The Early School Years: A Longitudinal Study. *Developmental Psychology, 28*, pp. 887–898.

Gentner, D. (1988). Metaphor As Structure Mapping: The Relational Shift. *Child Development, 59*, pp. 47–59.

Goodglass, H., and Geschwind, N. (1976). Language Disorders (Aphasia). In E. C. Carterette, and M. Friedman (Eds.), *Handbook Of Perception, Vol. 7*. New York: Academic Press.

Gopnik, M. (1990). Feature Blindness: A Case Study. *Language Acquisition, 1*, pp. 139–164.

Gopnik, M., And Crago, M. B. (1990). Familial Aggregation Of A Developmental Language Disorder. *Cognition, 39*, pp. 1–50.

Greenberg, J. (Ed.) (1978). *Universals Of Human Language, Vols. 1 To 4*. Stanford, CA: Stanford University Press.

Gupta, P., and MacWhinney, B. (1994). Is The Articulatory Loop Articulatory Or Auditory? Re–Examining The Effects Of Concurrent Articulation On Immediate Serial Recall. *Journal Of Memory And Language*, 33. pp. 63-88.

Gupta, P., and MacWhinney, B. (1995). Vocabulary Acquisition And Phonological Memory. *Brain And Language*. (in press).

Harrington, M. (1992). Second–Language Acquisition And Short–Term Memory. In R. Harris (Ed.). *Cognitive Processing In Bilinguals*. Amsterdam: Elsevier.

Ijaz, H. (1986). Linguistic And Cognitive Determinants Of Lexical Acquisition In A Second Language. *Language Learning, 36*, pp. 401–451.

Imedadze, N., And Tuite, K. (1992). The Acquisition Of Georgian. In D. I. Slobin (Ed.), *The Crosslinguistic Study Of Language Acquisition: Volume 3*. Hillsdale, NJ: Lawrence Erlbaum.

Johnson, J., And Newport, E. (1989). Critical Period Effects In Second–Language Learning: The Influence Of Maturational State On The Acquisition Of English As A Second Language. *Cognitive Psychology, 21*, pp. 60–99.

Johnson, J. S., And Newport, E. L. (1991). Critical Period Effects On Universal Properties Of Language: The Status Of Subjacency In The Acquisition Of A Second Language. *Cognition, 39*. pp. 215–258.

Kail, M. (1989). Cue Validity, Cue Cost, And Processing Types In French Sentence Comprehension. In B. MacWhinney, And E. Bates (Eds.), *The Crosslinguistic Study Of Language Processing*. New York: Cambridge University Press.

Kail, R. (1992). Processing Speed, Speech Rate, And Memory. *Developmental Psychology, 28*, pp. 899–904.

Kawamoto, A. (1993). Non–Linear Dynamics In The Resolution Of Lexical Ambiguity: A Parallel Distributed Processing Account. *Journal Of Memory And Language, 32*, pp. 474–516.

Kawamoto, A., And Zemblidge, J. (1992). Pronunciation Of Homographs. *Journal Of Memory And Language, 31*, pp. 349–374.

Keenan, J., And MacWhinney, B . (1987) . Understanding The Relation Between Comprehension And Production. In H. W. Dechert, And M. Raupach (Eds.), *Psycholinguistic Models Of Production*. Norwood, N.J.: ABLEX.

Kempe, V., And MacWhinney, B. (1994). The Crosslinguistic Assessment of Foreign Language Vocabulary. *Applied Psycholinguistics, 17*, (in press).

Kilborn, K. (1987). *Sentence Processing In A Second Language: Seeking A Performance Definition Of Fluency*. University Of California At San Diego.

Kilborn, K. (1989). Sentence Processing In A Second Language: The Timing Of Transfer. *Language And Speech, 32*, pp. 1–23.

Kintsch, W. (1977). On Comprehending Stories. In M. Just, And P. Carpenter (Eds.), *Cognitive Processes In Comprehension*. New York: Wiley.

Krashen, S. (1978). Individual Variation In The Use Of The Monitor. In W. Ritchie (Ed.), *Principles Of Second–Language Learning*. Academic Press: New York, NY.

Krashen, S. (1982). *Principles And Practice In Second–Language Acquisition*. New York: Pergamon Press.

Kroll, J. (1990). Recognizing Words And Pictures In Sentence Contexts: A Test Of Lexical Modularity. *Journal Of Experimental Psychology: Learning, Memory, And Cognition, 16*, pp. 747–759.

Kroll, J., And Borning, L. (1987). Shifting Language Representations In Novice Bilinguals: Evidence From Sentence Priming . *Unpublished Manuscript, Talk Presented At The 27th Annual Meeting Of The Psychonomic Society*.

Kroll, J., And Potter, M. (1984). Recognizing Words, Pictures, And Concepts: A Comparison Of Lexical, Object, And Reality Decisions. *Journal Of Verbal Learning And Verbal Behavior. 23*, pp. 39–66.

Kroll, J., And Sholl. (1992). Lexical and Conceptual Memory in Fluent and Nonfluent Bilinguals. In Harris, R. (Ed.). *Cognitive Processing in Bilinguals*. Amsterdam: North Holland.

Lado, R. (1971). Second–Language Teaching. In C. E. Reed (Ed.)'(Eds.), The Learning Of Language . New York: Appleton–Century–Crofts.

Lightfoot D. (1989) The Child's Trigger Experience. Degree-o Learnability. *Behavioral and Brain Sciences, 12*, pp. 221-275

Liljencrants, J., And Lindblom, B. (1972). Numerical Simulation Of Vowel Quality Systems: The Role Of Perceptual Contrast. *Language, 4*, pp. 839–862.

Lively, S., Pisoni, D., And Logan, J. (1990). Some Effects Of Training Japanese Listeners To Identify English Ir/ And 11/. In Y. Tohkura (Ed.), *Speech Perception, Production And Linguistic Structure*. Tokyo: OHM.

Lombardi, L., And Potter, M. (1992). The Regeneration Of Syntax In Short Term Memory. *Journal Of Memory And Language, 31*, pp. 713–733.

MacNeilage, P. F. (1970). Motor Control Of Serial Ordering Of Speech. *Psychological Review, 77*, pp. 182– 196.

MacWhinney, B. (1978). The Acquisition Of Morphophonology. *Monographs Of The Society For Research In Child Development, 43*.

MacWhinney, B. (1987). The Competition Model. In B. Macwhinney (Ed.), *Mechanisms Of Language Acquisition*. Hillsdale, NJ: Lawrence Erlbaum.

MacWhinney, B. (1992). Transfer And Competition In Second–Language Learning. In R. Harris (Ed.), *Cognitive Processing In Bilinguals*. Amsterdam: Elsevier.

MacWhinney B. (1994) The Computational analysis of Interactions. In Fletcher, P. and B. MacWhinney (Eds.), *The Handbook of Child Language*. Oxford: Blackwell.

MacWhinney, B., And Bates, E. (Eds.). (1989). *The Crosslinguistic Study Of Sentence Processing*. New York: Cambridge University Press.

MacWhinney, B., Leinbach, J., Taraban, R., And McDonald, J. (1989). Language Learning: Cues Or Rules? *Journal Of Memory And Language, 28*, pp. 255–277.

McDonald, J., And MacWhinney, B. (1989). Maximum Likelihood Models For Sentence Processing Research. In B. Macwhinney, And E. Bates (Eds.), *The Crosslinguistic Study Of Sentence Processing*. New York: Cambridge University Press.

McDonald, J., And MacWhinney, B. (1994). The Time Course Of Anaphor Resolution: Effects Of Implicit Verb Causality And Gender. *Journal Of Memory And Language*, 33.

McNeil, M., And Kent, R. (1990). Motoric Characteristics Of Adult Aphasic And Apraxic Speakers. In G. E. Hammond (Ed.), *Cerebral Control Of Speech And Limb Movements* Amsterdam: Elsevier.

Mills, A. (1986). The Acquisition Of German. In D. I. Slobin (Ed.), *The Crosslinguistic Study Of Language Development*. Hillsdale, N. J.: Lawrence Erlbaum Associates.

Naiman, N., Frohlich, M., Stern, H. H., And Todesco, A. (1978). *The Good Language Learner*. Toronto: Ontario Institute For Studies In Education.

Odell, K., McNeil, M., Rosenbek, J., And Hunter, L. (1991). Perceptual Characteristics Of Vowel And Prosody Production In Apraxic, Aphasic, And Dysarthric Speakers. *Journal Of Speech And Hearing Research, 34*, pp. 67–80.

O'Malley, M., And Chamot, A. (1990). *Learning Strategies In Second–Language Acquisition*. Cambridge: Cambridge University Press.

Oyama, S. (1976). A Sensitive Period For The Acquisition Of A Nonnative Phonological System. *Journal Of Psycholinguistics Research, 5*, pp. 261–283.

Papagno, C., Valentine, T., And Baddeley, A. (L991). Phonological Short–Term Memory And Foreign–Language Vocabulary Learning. *Journal Of Memory And Language, 30*, pp. 331–347.

Perfetti, C. A., Bell, L. C., And Delaney, S. M. (1988). Automatic (Prelexical) Phonetic Activation In Silent Word Reading: Evidence From Backward Masking. *Journal Of Memory And Language, 27*, pp. 59–70.

Pinker, S. (1989). *Learnability And Cognition: The Acquisition Of Argument Structure*. Cambridge: MIT Press.

Pinker, S. (1991). Rules Of Language. *Science, 253*, pp. 530–535.

Plaut, D., And McClelland, J. (1993). Generalization With Componential Attractors: Word And Nonword Reading In An Attractor Network. In *Proceedings Of The Fifteenth Annual Conference Of The Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Plaut, D., And Shallice, T. (1991). Effects Of Word Abstractness In A Connectionist Model Of Deep Dyslexia. In *Proceedings Of The Thirteenth Annual Conference Of The Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Schmajuk, N., And DiCarlo, J. (1992). Stimulus Configuration, Classical Conditioning, And Hippocampal Function. *Psychological Review, 99*, pp. 268–305.

Seidenberg, M., And McClelland, J. (1989). A Distributed, Developmental Model Of Word Recognition And Naming. *Psychological Review, 96*, pp. 523–568.

Squire, L. R. (1992). Memory And The Hippocampus: A Synthesis From Findings With Rats, Monkeys, And Humans. *Psychological Review, 99*, pp. 195–231.

Squire, L. (1992). Declarative And Nondeclarative Memory: Multiple Brain Systems Supporting Learning And Memory. *Journal Of Cognitive Neuroscience, 4*, pp. 233–243.

Tallal, P., And R., S. (1980). Speech Perception Of Language Delayed Children. In G. H. Yeni–Komshian, J. F. Kavanagh, And C. A. Ferguson (Eds.), *Child Phonology: Perception*. New York: Academic Press.

Talmy, L. (1976). Semantic Causative Types. In M. Shibatani (Ed.), *Syntax And Semantics*. New York: Academic Press.

Talmy, L. (1977). Rubber–Sheet Cognition In Language. In S. F. W. Beach, And S. Philosoph (Eds.), *Papers From The Thirteenth Regional Meeting*. Chicago: Chicago Linguistic Society.

Talmy, L. (1988). Force Dynamics In Language And Cognition. *Cognitive Science, 12*, pp. 59-100.

Trahey, M., And White, L. (1993). Positive Evidence And Preemption In The Second-Language Classroom. *Studies In Second Language Acquisition, 15*, pp. 181–204.

Urosevic, Z., Carello, C., Lukatela, G., Savic, M., And Turvey, M. T. (1988). Word Order And Inflectional Strategies In *Syntactic Processing. Language And Cognitive Processes, 31*, pp. 49–71.

Van Der Lely, H. (1993). Canonical Linking Rules: Forward Vs. Reverse Linking Normally Developing And Specifically Language Impaired Children. *Cognition*, 51, pp. 29-72.

Van Der Lely, H., And Howard, D. (1993). Children With Specific Language Impairment: Linguistic Impairment Of Short-Term Memory Deficit. *Journal Of Speech And Hearing Research*, 36, pp. 34-82.

VanDijk, T., And Kintsch, W. (1983). *Strategies Of Discourse Comprehension*. New York: Academic Press

Venezky, R. (1970). *The Structure Of English Orthography*. The Hague: Mouton.

Werker, J. F., Gilbert, J. H. V., Humphrey, K., And Tees, R. C. (1981). Developmental Aspects Of Cross-Language Speech Perception. *Child Development, 52*, pp. 349–355.

White, L. (1989). The Adjacency Condition On Case Assignment: Do Learners Observe The Subset Principle? In S. Gass, And J. Schachter (Eds.), *Linguistic Perspectives On Second-Language Acquisition*. Cambridge: Cambridge University Press.

White, L. (1990). The Verb–Movement Parameter In Second–Language Acquisition. *Language Acquisition, 4*, pp. 337–360.

White, L. (1991). Adverb Placement In Second–Language Acquisition: Some Effects Of Positive Negative Evidence In The Classroom. *Second Language Research, 7*, pp. 133–161.

White, L. (1992). Long And Short Verb Movement In Second–Language Acquisition. *Canadian Journal Of Linguistics*, 37, pp. 273–286.

TEST THEORY AND LANGUAGE LEARNING ASSESSMENT[1]

Robert J. Mislevy
Educational Testing Service

HOLMES: *In solving a problem of this sort, the grand thing is to be able to reason backward. That is a very useful accomplishment but people do not use it much. In everyday affairs of life it is more useful to reason forward, and so the other comes to be neglected. There are fifty who can reason synthetically for one who can reason analytically.*

WATSON: *I confess I do not follow you.*

HOLMES: *I hardly expected that you would. Let me see if I can make it clearer. Most people, if you describe a train of events to them will tell you what the results would be. They can put those events together in their mind, and argue from them that something will come to pass. There are few, however, who, if you told them a result, would be able to evolve from their own inner consciousness what the steps were which led up to that result. This power is what I mean when I talk about reasoning backward, or analytically.*

WATSON: *I understand. (Doyle, 1930, p. 268).*

Introduction

Test theory, as we usually think of it, is part of a package. It encompasses models and methods for drawing inferences about what students know and can do—as cast in a particular framework of ideas from measurement, education, and psychology that coalesced in the first third of the twentieth century. In a nutshell, (i) human abilities were viewed as traits, or 'relatively stable characteristics of a person—attributes, enduring processes, or dispositions—which are consistently manifested to some degree when relevant, despite considerable variation in the range of settings and circumstances' (Messick, 1989, p. 15); (ii) traits were conceived as numbers along measurement scales, locating people along continua of mental characteristics just as their heights and weights located them along continua of physical characteristics; (iii) tendencies in behavior in samples of a domain of discrete settings and circumstances (e.g., assessment tasks) were the privileged form of evidence about traits; and (iv) the purpose of test theory was to guide reasoning from observed behavior in samples of situations from the domain to inferences about traits.

This 'domain–behavior' framework of assessment generates a universe of discourse: the nature of the problems one perceives, the kinds of statements one makes about students, the ways one gathers data to support them. Test theory, as we usually think of it, is the

---

[1] In addition to being presented as a plenary address at the CALL 1994 Language Aptitude Invitational Symposium, this paper also appeared in Language Testing 12 (3), pp. 341-369. I am grateful for discussions with Nancy Anderson, Dan Eignor, Anne Harvey, and Ming Mae Wang.

application of inferential principles to deal with such problems as missing data, source unreliability, multi-stage inference, conflicting or overlapping observations, multiple sources of disparate evidence, and constrained resources for gathering and evaluating information—as they arise in this framework.

The views of the nature and the acquisition of competence in a second language, and the nature of inferences we would wish to make about students' developing competence, do not always fall within this familiar realm. In particular, we may wish to wish to take into account the situated and contextual aspects of language learning, and we may wish to gather data from complex tasks that stress the inter-connections among aspects of students' competence. But while these developments may suggest student models and observational strategies quite different from those employed by Spearman, Thurstone, and Thorndike, practical work under alternative perspectives inevitably faces in some form the same general inferential problems listed above. This is where a more broadly construed conception of test theory is required. It is not sufficient merely to define the class of conjectures about student competence we wish to address, and devise settings in which students can display these competencies. We must, further, specify how what we observe is related to competence as we choose to conceive it, and construct a framework for carrying out inference within the framework we thus erect.

To this end, the following section discuss the notions of evidence and inference more broadly than they are usually conceived in educational assessment. The role of probability-based inference in assessment is described. Ideas are then illustrated with two language-learning assessment challenges—contextual effects on learning and complex performance tasks—with regard to inference in the conceptual framework of the American Council of Teachers of Foreign Languages (ACTFL) guidelines (ACTFL, 1989).

Evidence and Inference

Inference is reasoning from what we know and what we observe to explanations, conclusions, or predictions. The skills we must apply in educational assessment are essentially the same as those employed in such fields as troubleshooting, medical diagnosis, criminology, and intelligence analysis. We attempt to establish the weight and coverage of evidence in what we observe. The very first question we must address is "Evidence about what?" Schum (1987, p. 16) points out the crucial distinction between *data* and *evidence*: "A datum becomes evidence in some analytic problem when its *relevance* to one or more hypotheses being considered is established. . . . [E]vidence is relevant on some hypothesis if it either increases or decreases the likeliness of the hypothesis. Without hypotheses, the relevance of no datum could be established."

Test data acquire meaning only in relation to particular hypotheses, or conjectures, that we entertain. The same observation can be direct evidence for some conjectures and indirect evidence for others, and wholly irrelevant to still others. In educational assessment, we construct our conjectures around notions about the nature and the acquisition of competence. We can actually observe only the specific actions and products that students

produce in specific circumstances. To evaluate their progress or guide further instruction, however, we talk at a higher level of abstraction, using specific observations as evidence for our inferences.

A conception of competence is effected as a set of variables in a student model, a simplified description of selected aspects of the infinite varieties of skills and knowledge that characterize real students. Depending on our purposes, we might distinguish anywhere from one or hundreds of facets. They might be expressed in terms of numbers, categories, or some mixture; they might be conceived as persisting over long periods of time, or apt to change at the next problem-step. They might concern tendencies in behavior, conceptions of phenomena, available strategies, or levels of development. The point is that we don't observe these variables directly. We observe only student's behavior in limited circumstances—indirect evidence about competence more abstractly conceived. Test theory, broadly construed, is conceptual and statistical machinery for reasoning from observations to inferences in terms of the competence model.

Suppose we want to make a statement about Jasmine's proficiency, in terms of likely values of the variables in a model built around some key aspects of competence. We can't observe these values directly,[2] but perhaps we can make an observation that bears information about the plausibility of various values under the model: her answer to a multiple-choice question, say, or two sets of judges' ratings of her violin solo, or an essay outlining how to determine which paper towel is most absorbent. The observation can't tell us her value with certainty, because similar behavior could be produced by students with different underlying levels of competency depending on factors such as their familiarity with the context and situation. It is, however, more likely to be produced by students at some levels than others. Nonsensically answering '¿Como está usted?' with 'Me llamo Carlos,' for example, is much more likely from a student classified as a Low-Novice under the American Council of Teachers of Foreign Languages (ACTFL) Reading guidelines (see Table 1) than an Advanced student. (There are similarly conceived guidelines for Writing, Speaking, and Listening.)

Conceptions of Competence

A conception of student competence and a purpose for assessment should drive the particular methods we need to get students to act in ways that reveal something about their competencies, or the forms of assessment we employ. This section contrasts key aspects of two broadly cast assessment paradigms, which we shall refer to as the 'domain-behavior' and 'cognitive/developmental,' paradigms, and notes some implications for assessment forms and test theory.

---

[2] After all, the model itself isn't truth but a simplified approximation we have constructed, and variable values are not so much characteristics of Jasmine, but of summaries of our knowledge about patterns we perceive in Jasmine's behavior, as seen through the lens of the model.

The 'domain–behavior' paradigm originated under trait psychology and evolved further under behaviorist psychology. From trait psychology came the notions of characterizing characteristics of persons in terms of numbers on a measurement scale, and taking as evidence for these numbers, counts of keyed behaviors in samples from a domain of relevant settings (such as test items). The following quotation reflects how this perspective came to be applied to the development and practice of educational assessment:

> *The educational process consists of providing a series of environments that permit the student to learn new behaviors or modify or eliminate existing behaviors and to practice these behaviors to the point that he displays them at some reasonably satisfactory level of competence and regularity under appropriate circumstances. The statement of objectives becomes the description of behaviors that the student is expected to display with some regularity. The evaluation of the success of instruction and of the student's learning becomes a matter of placing the student in a sample of situations in which the different learned behaviors may appropriately occur and noting the frequency and accuracy with which they do occur* (Krathwohl and Payne, 1971, p. 17–18).

Under the domain–behavior approach, the specification of an assessment describes a collection of task contexts as seen from the assessor's point of view, and provides a system for classifying the responses students might make. Potential responses in some contexts, such as multiple–choice items, are unambiguously right or wrong; in others, counts or instances of behaviors of certain types, the distinction of which may require expert judgment, are recorded. Behavior observed in a sample of tasks constitutes direct evidence for expected behavior in the domain as a whole, which in turn constitutes an operational definition of competence. The primary inferential task of standard test theory is to characterize the weight of evidence that samples of tasks provide about students' domain proficiencies. The processes by which students *acquire* competence are of interest, of course, to students, teachers, and researchers alike, but for the most part, these questions lie outside the universe of discourse associated with the domain–behavior paradigm of assessment (Stake; 1991).

In contrast, the acquisition of competence plays a central role in contemporary cognitive and educational psychology. The following quotation reflects the cognitive/developmental perspective as it relates to educational assessment:

> *Essential characteristics of proficient performance have been described in various domains and provide useful indices for assessment. We know that, at specific stages of learning, there exist different integrations of knowledge, different forms of skill, differences in access to knowledge, and differences in the efficiency of performance. These stages can define criteria for test design. We can now propose a set of candidate dimensions along which subject matter competence can be assessed. As competence in a subject matter grows, evidence of a knowledge base that*

> *is increasingly* coherent, principled, useful, *and* goal-oriented *is displayed, and test items can be designed to capture such evidence* (Glaser, 1991, p. 26, emphasis in original).

From the cognitive perspective, the specifications for an assessment describe contexts that can evoke evidence about students' competence as conceived at a higher level of abstraction, and provide judgmental guidelines for mapping from observed behavior to this inferred competence. This behavior provides evidence about competence so conceived, but not necessarily *direct* evidence. We may have to interpret this behavior in light of additional knowledge or supporting evidence about, for example, how the content or the context of a task interacts with the student; we may need to infer, or learn more about, the task as seen from the point of view of the student.

The ACTFL reading proficiency guidelines (Table 1) illustrate this point. Contrast the description of Intermediate readers' competence with texts 'about which the reader has personal interest or knowledge' with Advanced readers' competence with '...texts which treat unfamiliar topics and situations.' This distinction is fundamental to the underlying conception of developing language proficiency, but obviously a situation that is familiar to one student is unfamiliar to others. The evidential import of the same behavior in the same situation can differ radically for different students, and, as we shall explore further, affect what we infer about their capabilities from their behavior.

Probability-Based Inference

> *Probability isn't really about numbers; it's about the structure of reasoning.*
> Glenn Shafer (quoted in Pearl, 1988, pp. 44)

As the preceding section addressed what we want to reason about in educational assessment, this section concerns how we want to reason. It outlines the basic kinds of reasoning tasks we face, and reviews some tools from probability theory we can gainfully employ to this end, some hundreds of years old and others quite recent.

Kinds of Inference

Schum (*1987*) distinguishes among deductive, inductive, and abductive reasoning, all of which play essential and interlocking roles in educational assessment:

- *Deductive reasoning* flows from generals to particulars, within an established framework of relationships among variables—from causes to effects, from diseases to symptoms, from the way a crime is committed to the evidence likely to be found at the scene, from a student's knowledge and skills to observable behavior. Under a given state of affairs, what are the likely outcomes?

- *Inductive reasoning* flows in the opposite direction, also within an established framework of relationships—from effects to possible causes, from symptoms to

possible diseases, from a student's solution to likely configurations of knowledge and skill. Given the outcomes we see, what state of affairs may have produced them?

- *Abductive reasoning* (a term coined by the philosopher Charles C. Peirce) proceeds from observations to new hypotheses, new variables, or new relationships among variables. 'Such a 'bottom-up' process certainly appears similar to induction; but there is an argument that such reasoning is, in fact, different from induction since an existing hypothesis collection is enlarged in the process. Relevant evidentiary tests of this new hypothesis are then *deductively* inferred from the new hypothesis.' (Schum, *1987*, p. 20).

Conjectures, and the understanding of what constitutes evidence about them, emanate from the variables, concepts, and relationships of the field within which reasoning is taking place. The theories and explanations of a field suggest the structure through which deductive reasoning flows—the 'generative principles of the domain,' to borrow a phrase from Greeno (1989). Inductive and abductive reasoning depend just as critically on the same structures, as the task is to speculate on circumstances which, when their consequences are projected deductively, lead plausibly to the evidence at hand. Determining promising possibilities, we reason deductively to other likely consequences—potential sources of corroborating or disconfirming evidence for our conjectures.

A detective at the scene of a crime reasons abductively to reconstruct the essentials and principals of the event. Anything he sees, in light of a career of experience, can suggest possibilities; ways things might happened which, reasoning deductively, could have produced the present state of affairs (e.g., documents, testimony, physical evidence). Given tentative hypotheses, does inductive reasoning from other observations conflict or fit in? When they conflict, does their juxtaposition spark a new hypothesis? A successful investigation leads to a plausible explanation of the case, which, reasoning deductively, supports the data at hand.

Mathematical Probability

Given key concepts and relationships, inferential objectives, and data, how should reasoning proceed? How can we characterize the nature and force of persuasion a mass of data conveys about a target inference? Workers in every field have had to address these questions as they arise with the kinds of inference and the kinds of evidence they normally address. Historically, the quest for principles of inference at a level that might transcend the particulars of fields and problems has received most attention in the fields of probability and statistics (unsurprisingly), philosophy, and jurisprudence. Our interest is in the first of these, and, in particular, mathematical or Pascalian (after Blaise Pascal) probability. For our purposes, the essential elements are a specified space of outcomes, or sample space; a parameter space; and a function that specifies the probabilities of outcomes given parameters, where probabilities are numbers between 0 and 1 that correspond to strength of belief and follow a few simple rules of combination for 'events,' where a 'Pascalian event' is a subset of the sample space. It is portentous that given

parameter values, we can express the relative likeliness of a Pascalian event as compared to any other events, and given an event, we can express the relative likeliness of a given parameter value as compared to any other parameter value.

When it is possible to map the salient elements of an inferential problem into the framework of mathematical probability, powerful tools become available to combine explicitly the evidence that various probans (elements of evidence or intermediate conjectures) convey about probanda (target conjectures), as to both weight and direction of probative force. Inferential subtleties such as chains of inferences, missingness, disparateness of sources of evidence, and complexities of inter-relationships among probans and probanda, can be resolved. A properly-structured statistical model embodies the salient qualitative patterns in the application at hand, and spells out, within that framework, the relationship between conjectures and evidence. It overlays a substantive model for the situation with a model for our knowledge of the situation, so that we may characterize and communicate what we come to believe—as to both content and conviction—and why we believe it—as to our assumptions, our conjectures, our evidence, and the structure of our reasoning.

Perhaps the two most important building blocks are conditional independence and Bayes theorem. Conditional independence is a tool for mapping Greeno's 'generative principles of a domain' into the framework of mathematical probability, expressing the substantive theory upon which deductive reasoning in a field is, and must be, based. This accomplished, Bayes theorem is a tool for reversing the flow of reasoning—inductively, from observations to the more fundamental concepts of the domain, through these same structures, to expressions of revised belief in the language of mathematical probability.

Conditional Independence

Two random variables x and y are *independent* if their joint probability distribution $p(x,y)$ is simply the product of their individual distributions—$p(x,y) = p(x)p(y)$. These variables are unrelated, in the sense that knowing the value of one provides no information about what the value of the other might be. Conditionally independent variables seem to be related $p(x,y) \neq p(x)p(y)$—but their co-occurrence can be understood as determined by the values of one or more other variables $p(x,y\backslash z) = p(x\backslash z)p(y\backslash z)$—, where the conditional probability distribution (px/z) is the distribution of values of x, given the value z of another variable. The conjunction of sneezing, watery eyes, and a runny nose described as a 'histemic reaction' could be triggered by various causes such as an allergy or a cold; the specific symptoms play the role of x's and y's, while the status of reaction-causing conditions plays the role of z. The paradigms of a field supply 'explanations' of phenomena in terms of concepts, variables, and putative conditional independence relationships. Judah Pearl (1988:44) argues that inventing intervening variables is not merely a technical convenience, but a natural element in human reasoning:

> *[C]onditional independence is not a grace of nature for which we must wait passively, but rather a psychological necessity which we satisfy*

> *actively by organizing our knowledge in a specific way. An important tool in such organization is the identification of intermediate variables that induce conditional independence among observables; if such variables are not in our vocabulary, we create them. In medical diagnosis, for instance, when some symptoms directly influence one another, the medical profession invents a name for that interaction (e.g., 'syndrome,' 'complication,' 'pathological state') and treats it as a new auxiliary variable that induces conditional independence; dependency between any two interacting systems is fully attributed to the dependencies of each on the auxiliary variable.*

In educational assessment, the variables in the student–competence model play the role of explanatory variables. They constitute the more abstract space in which we attempt to understand students' actions, evaluate their developing competences, and plan further instruction. From the point of view of mathematical probability, the starting point for assessment is deductive reasoning through such a framework: 'how likely is a particular observation, from each of the possible values in the competence model?' The answer—the 'likelihood function' induced by this particular possible response—conveys the information that the observation conveys about competence, in the way competence is being conceived. If the observation is equally likely from students at all values of the variables in the competence model, it carries no information for inferences about those variables. If it is likely at some values but not others, it sways our belief in those directions, with strength in proportion to how much more likely the observation is at those values.

To illustrate this deductive stage of reasoning, we will use a student model based on the ACTFL reading guidelines. We will work with three collapsed levels of reading proficiency, namely, novice, intermediate, and advanced, and map out the evidential grounding of two reading tasks, a multiple-choice question that is simply right or wrong and an extended performance task that supports four distinguishable levels of performance. We will assume for the moment that the requirements of background knowledge can be neglected. (This is *not* the case in many performance assessment tasks, and we shall discuss how to extend the framework to deal with this in the following section below on 'contextual dependencies.')

For each of the four reading competence categories, a panel in Figure 1 shows the probabilities of the different possible performance levels on the extended task. Each rectangle is a variable, with the probabilities associated with its different possibilities represented by bars that add up to one. Dashed bars represent certain knowledge—in Figure 1, looking at probabilistic expectations of responses if student competence level were known for a fact. The directed arrow in this so-called 'directed acyclic graph' (DAG) indicates the flow of deductive reasoning. We see that students at higher ACTFL levels are increasingly likely to do well on this task, although there is some chance for even advanced students to fare poorly and for novices to score well; that is, even knowing ACTFL with certainty would not give us perfect predictions of response. This is

reasoning *from* an abstract conception of competence *to* expected performance—the 'forward reasoning' Holmes described to Watson. We determine these probabilities through theory, expert judgment, model-fitting (e.g., a latent class or item response theory model), empirical data-gathering (e.g., observations on groups of students ascertained from external information to function at each of the three levels), or some combination (Andreassen, et al., 1987, illustrate these considerations in the context of medical diagnosis). Figure 2 shows similar conditional probabilities for the multiple-choice task. This hypothetical item is relatively easy, so we see in Figure 2 that only the novices will probably miss it. Intermediate students have 85% chances of getting it right and advanced students have 95% chances.

Bayes Theorem

We must reason inductively in most practical applications. In the language task example, we will observe a student's performances in order to increase our knowledge about a student's level of competence on the ACTFL scale. When we can satisfactorily explicate the probabilities of observations given (inherently unobservable) values of variables in the student model as was illustrated above, Bayes theorem provides a mechanism for reversing the flow of reasoning in a coherent manner. The mathematics of Bayes Theorem can be found in any statistical text; its central role in cognitive diagnosis and educational assessment is discussed more fully in Mislevy (1994,1995). The essential idea is as follows:

- Before seeing observations, our belief about possible values of variables in the student model is expressed as a probability distribution—the *prior distribution*.
- A particular value of an observable variable provides evidence about those values, in proportion to its probability of occurrence under each—the *likelihood function*.
- The product of the prior distribution and the likelihood function yield, for each possible value in the student model, a value proportional to its probability in a new distribution that reflects our revised beliefs—the *posterior distribution*.

Figure 3 represents inductive reasoning with the extended performance task. Inference flows in the opposite direction of the relationships represented by the directed arrow, which constitute the theory-driven structure of deductive reasoning—Holmes' 'backwards reasoning.' Now values of task performance become known with certainty when they are observed, and beliefs about possible values in the student model are updated. Each panels depicts the posterior probabilities for student competence induced by observing one of the four possible performance levels, starting from a prior distribution that considered the three levels equally likely. (In this special case, the posterior distribution is proportional to the likelihood function.) We see that, as would be expected, higher levels of observed performance shifts our beliefs about students toward higher levels of competence. Figure 4 shows similar results for the multiple-choice task. Because this item is easy, a wrong response shifts our belief sharply toward a student

being a novice, while a right response shifts belief away from novice, but does not provide much information to distinguish between intermediate and advanced.

Bayesian Inference Networks

Carrying out probability–based inference efficiently in complex networks of interdependent variables is an active topic in statistical research, spurred by applications in such diverse areas as forecasting, pedigree analysis, troubleshooting, and medical diagnosis. Interest centers on obtaining the distributions of selected variables conditional on observed values of other variables, such as likely characteristics of offspring of selected animals given characteristics of their ancestors, or probabilities of disease states given symptoms and test results. The conditional independence relationships suggested by substantive theory play a central role in the topology of the network of inter–relationships in a system of variables. If the topology is favorable, such calculations can be carried out efficiently through generalizations of Bayes theorem even in very large systems, by means of strictly local operations on small subsets of inter–related variables ('cliques') and their intersections. Discussions of construction and local computation in such Bayesian inference networks can be found in the statistical and expert–systems literature (see, for example, Lauritzen and Spiegelhalter, 1988, and Shafer and Shenoy, 1988; computer programs that carry out the required computations include Andersen, Jensen, Olesen, and Jensen, 1989, and Noetic Systems, 1991).

Figure 5 is a DAG for a simple inference network that combines the multiple–choice and extended–performance tasks introduced above. The three panels depict how belief about a student's level of competence is updated as the two responses are observed in turn. Directed arrows run from the student–model competence variable to each of the tasks, but there is no direct connection between the two; this indicates that they are conditionally independent given level of competence. It is in establishing such relationships that substantive theory comes into play: in defining unobservable variables that characterize students' state or structure of understanding, and observable variables that will convey evidence about that understanding; in defining intervening variables and conditional independences through which deductive reasoning flows, so as to capture important substantive relationships and simplify computations. Note again the distinction between those assessment variables that are potentially observable and 'student–model variables' that are not, but in terms of which theories of knowledge and learning are framed (Mislevy, 1995).

The following sections extend our running ACTFL example in two ways in order to illustrate inferences about language competence that take into account the role of context and background in language acquisition and of observing more complex performances that require multiple aspects of competence. The focus is on the way this knowledge about the kind of competence we wish to make inferences about, and the way that it is manifest in complex settings, can be dealt with using probability–based inference.

Dealing with Context and Situation

> *[I]t appears that research on the measurement of the intellectual abilities generally associates with the term intelligence reached a point of diminishing returns a number of decades ago; though there has been continuing refinement of technical methods for test construction, progress has remained essentially asymptotic with regard to problems of predicting intellectual functioning outside of testing situations. An important reason suggested by the present analysis is continuing overdependence on the concept of context-free ability tests and consequent lack of analysis of the interactions and contexts* (Estes, 1981, pp. 18–19).

The 'traits' that achievement tests purportedly measure, such as 'mathematical ability,' 'reading level,' or 'physics achievement,' do not exist *per se*. While test scores do tell us something about what students know and can do, any assessment task stimulates a unique constellation of knowledge, skill, strategies, and motivation within each examinee. To some extent in any assessment comprising multiple tasks, which ones are relatively hard for some students are relatively easy for others, depending on the degree to which the tasks relate to the' knowledge structures that students have, each in their own way, constructed. From the domain–behavior perspective, this is 'noise,' or measurement error. It obscures what one is interested in, namely, locating people along a single dimension as to a *general* behavioral tendency, and tasks that don't line up people in the same way are less informative than ones that do.

From the cognitive/developmental perspective, however, these interactions are fully expected, since knowledge typically develops first in context, then is extended and decontextualized so that it can be applied to more broadly to other contexts. A given task may thus have the potential of providing considerable information about a given student, or none at all. Standard test theory does not address this concern at the level of tasks, but at the level of the combined test scores only after averaging results over multiple tasks; this is the issue of 'test validity' (Messick,1989). But the greater investment each task requires and the more contextual knowledge it demands, the less efficient this approach becomes; hence the so-called 'low generalizability' problem some writers have attributed to performance assessments (e.g., Shavelson, Baxter, and Pine, 1992). The in-depth project on proportionality that provides solid assessment information and a meaningful learning experience for the students whose prior knowledge structures it dovetails, becomes an unconscionable waste of time for students for whom it has no connection. The alternative is to take contextual and/or situational data into account when determining the evidential value that tasks provide about'students' competencies. Practical assessment methods for doing this are discussed below. First, however, we illustrate the inferential situation with an extended inference network.

The mileposts outlined in the ACTFL reading guidelines are based on empirical evidence and theories about how competence in acquiring information from text in a foreign language develops. We have noted the contrast between intermediate readers'

competence with texts 'about which the reader has personal interest or knowledge' with advanced readers' comprehension of 'texts which treat unfamiliar topics and situation'—a distinction fundamental to the underlying conception of developing language proficiency, which can alter the evidential import of the same behavior from the two students about their ACTFL levels. These relationships can be incorporated into a Bayesian inference network by extending the structure beyond nodes that characterize the situation only from an 'objective' point of view that pertains equally to all students—to nodes that vary across students in connection with their particular points of view; for example, whether a student has read a book upon which a reading passage is based. Consider an inference network that extends the one shown in Figure 1 by adding a new contextual variable, namely, whether the student is familiar or unfamiliar with the book in question.

Figure 6 illustrates expectations about performance as a function of given values of competence level and context familiarity, or the by now familiar flow of deductive reasoning. Note the different expectations when the student is and is not familiar. Even students in the advanced category rarely perform well when they are unfamiliar with the context. When level of familiarity is not known, the expectations are an average of the two known conditions, and consequently much more diffuse. (The average is weighted by the proportion of students in each category who are and are not familiar with the book; for simplicity, this figure and the next assume a 50–50 split.) Figure 7 shows the results of inductive reasoning from observing a fairly low performance or a fairly high performance, under the conditions the we either (1) *know* the student is familiar, (2) *know* the student is *not* familiar, and (3) *don't know* whether the student is familiar. Note that the task conveys much more evidence about reading competence when we know the student is familiar with the context. That is, for a given level of observed performance, a more concentrated probability distribution, or a sharper inference, is obtained for level of proficiency if we know that the student is familiar with the context than if we know she is not, or if we don't know whether or not she is familiar. When low performance is observed in the third column where we don't know if the context is familiar to the student, appreciable probability remains that the student is intermediate or advanced; this is because both alternative explanations for low performance (low competence, and high competence but unfamiliar context) must be maintained.

Standard test theory for domain–behavior inferences faces the third situation illustrated above. There are two standard test–theory methods for handling context dependency interaction between students and tasks in a domain: minimize it as much as possible, then average over whatever interaction remains with as many tasks as feasible. Minimizing it is accomplished by using tasks with which all examinees are similarly familiar or similarly unfamiliar. The costs are (1) avoiding tasks with which students may be personally interested, acquainted, and able to display competences, and (2) making inferential errors of over– or underestimation of competence with respect to students for whom a particular task is atypically familiar or unfamiliar. Obviously the fewer tasks a student is administered, the more likely it is that this latter error occurs; therefore, averaging over as many tasks as possible helps to mitigate this problem. And it is an effective strategy with

short, distinct, tasks such as multiple-choice items. It is less effective as each task becomes more time consuming.

Two alternative ways of handling contextual and situational effects both attempt to move from the last column in Figure 7 to the first or second column—preferably the first because that is where evidential value is highest, but at least if you know you're in the second column, you can use this information appropriately! The first way is to obtain contextual and situational data from each student along with task performance data. To the extent possible, findings about background variables are entered in an inference network along with task responses, and the conditional relationships among background and performance are taken into account. This strategy is taken in large-scale educational surveys such as the International Assessments of Mathematics in the form of 'opportunity to learn' measures (Platt, 1975). It is not effective for assessing individuals because tasks are administered without regard to these effects. This is analogous to administering a large battery of unrelated diagnostic tests to a hospital patient before we have any idea what the problem is, then only later trying to sort out which ones were meaningful ('turns out he has a broken leg, so I guess we don't need any data from this CAT scan of his brain').

A second strategy is adapting what one observes to the student in accordance with values on what corresponds in our simple example to 'familiarity.' This can be done either by the assessor, as when an interviewer determines a subject of interest about which a conversation with a student can profitably take place, or by the student, as when choice among topics or exercises is provided. This is analogous in medical diagnosis to administering diagnostic tests sequentially, in light of previous results and improved conjectures, and to asking the patient to provide information about what hurts and what happened. The choice strategy for educational assessment is most likely to provide interpretable evidence of competence if, no matter what the choice, evidence must be provided about the same more generally described competence, and it is made clear to the examinee what it desired and how it will be evaluated. Myford and Mislevy (1995) and Mislevy (1995) discuss how this strategy is implemented and monitored in the College Entrance Examination Board's Advanced Placement Studio Art portfolio assessment.

Complex Interaction of Skills within Tasks

Resnick and Resnick (1989) argue persuasively against the decontextualized and decomposed assessment tasks that characterize standard achievement tests. Genuine expertise, they claim, is contextualized and calls upon multiple aspects of skill and knowledge in concert. If this is what we seek to develop in students, should not they learn and be assessed in like terms to a far greater than they typically are? Creating assessment tasks that tap meaningful learning in engaging and effective ways is a significant challenge, but there are signs of progress (see, e.g., Lesh and Lamon, 1992). There has been less progress in figuring out just what to do with the 'data' that one obtains when students perform the tasks, both as to identifying just what is meaningful and how the tasks are to be evaluated, and as to combining results across multiple and diverse tasks. This section

addresses the latter problem in the framework of Bayesian inference networks; the former problem is discussed, among other places, in Myford and Mislevy (1995).

Consider again the ACTFL guidelines for reading, writing, speaking, and listening. Suppose we want to assess students' competencies in Spanish in terms of these guidelines by means of the four tasks listed below. Figure 8 depicts the structure of the evidential relationships, showing baseline proportions of competence–levels and task performances in a population of interest—our state of knowledge about a student from this population before we see any of his or her performances. The connections among the aspects of competence reflect the possibility of empirical relationships among them in a population of interest (e.g., people who can write well in a foreign language might usually read well; a weaker relationship may exist between writing and listening).

- *Task A* is the extended performance reading task introduced above, providing a bit of direct evidence about reading only.[3] The relationship between Task A and Reading Competence is the one shown in Figure 1, but now embedded in a larger context.

- *Task B* is reading a complex passage and writing a response to a question about it. It is possible to obtain evidence about both reading and writing, but a dependency must be accounted for: low levels of writing competence eliminate the chance to acquire direct evidence about reading. A sensible response competently written provides evidence about higher competence about both reading and writing (the first panel of Figure 9). A well–written but off–task response shifts belief toward higher competence in writing but lower levels of competence in reading (the second panel of Figure 9). A poorly–written and off–target response shifts belief away from higher levels of both reading and writing (the final panel of Figure 9).

- *Task C* asks the student to listen to a taped conversation with a transcript provided, then talk about the interaction. A well–spoken and accurate response signifies higher speaking competence (see the first panel of Figure 10), and shifts beliefs about both listening and speaking higher—though not for either as much as for speaking, since we don't know whether the student listened to the conversation, read the transcript, or both. An 'okay' response shifts beliefs about speaking toward intermediate, and both listening and reading in the same direction—though again not as strongly

[3] Direct evidence about reading competence may provide *in*direct evidence about other competencies, to the extent that people who tend to do well in one aspect of language competence tend to do well in others. But the four–aspect ACTFL guidelines already embody the results' research on this topic: there are more finely detailed aspects of competence within reading that *do* tend to develop together, and are thus subsumed in the more generally defined reading guidelines; the same holds for listening, writing, and speaking. This finer breakdown would in fact be required in instruction. Competencies in the four main aspects, however, are seen to follow very different paths in different people. Graduate students may be required to learn to read a foreign language, for example, but acquire few listening or speaking skills. Conversely, extended visitors to a foreign country may pick up speaking and listening skills rapidly with only reading or writing skills.

because of the multiple explanations for this observation (the second panel of Figure 10). A 'poor' response shifts belief about all three aspects of competence involved in the tasks downward. Possible causes, the situations of which are averaged over in the result, include failure at the stage of understanding the message—i.e., lack of both listening and reading skills—and/or the stage of responding—i.e., low speaking skills (the final panel of Figure 10).

- *Task D* asks the student to listen to a taped conversation, and indicate by raising her hand when a business transaction is completed. Direct evidence about only listening competence is obtained. Figure 11 shows the results of observing a student respond correctly to Task D and do well on Task A after having done poorly on Task C. That is, the final panel of Figure 10 was the state of belief before observing this new correct response to Task D. Obtaining evidence that the student may have both reading and listening helps sort out the possibilities that could have led to poor performance in Task C; it is now more likely that speaking competence was the source of difficulty there.

For the reasons discussed above, I do not generally favor having holistic quality standards applied uniquely to individual tasks, each of which probes different mixtures of aspects of competence. The combination of idiosyncratic scores by any such means cannot capture differences among configurations of competence, and ignores patterns of strength or weakness among aspects of competence across tasks. The meaning of combined idiosyncratic scores is unambiguous only when almost all performances are successful or almost all are unsuccessful. I much prefer a structure under which evidence about various aspects of competence evinced by a task are evaluated in light of their mixture, accounting for their interdependencies. Having coherently interpreted evidence about aspects of competencies, one can then collapse this information in various ways for summarization, reporting, and evaluation. (See Haertel, 1989, and Haertel and Wiley, 1993, on the topic of explicating evidential structure of performance tasks.)

Conclusion

We do not build probability models for most of the reasoning we do, either in our jobs or our everyday lives. We continually reason deductively, inductively, and abductively, to be sure, but not through explicit formal models. Why not? Partly because we use heuristics, which, though suboptimal (e.g., Kahneman, Slovic, and Tversky, 1982), generally suffice for our purposes; more importantly, because much of our reasoning concerns domains we know something about. Attending to the right features of a situation and reasoning through the right relationships, informally or even unconsciously, provides some robustness against suboptimal use of available information within that structure. Heuristics, habits, rules of thumb, standards of proof, and typical operating procedures guide practice in substantive domains, more or less in response to what seems to have worked in past and what seems to have led to trouble. This inferential machinery co-evolves with, and is intimately intertwined with, the problems, the concepts, the constraints, and the methodologies of the field (Kuhn, 1970, p. 109). But difficulties arise

when inferential problems become so complex that the usual heuristics fail, when the costs of unexamined standard practices become exorbitant, or when novel problems appear. It is in these situations that more generally framed and formally developed systems of inference provide their greatest value.

We face this situation today in language learning assessment; indeed, in educational assessment in general. The standard methods, rules of thumb, and canons of good practice have evolved to address inference in a universe of discourse more restricted with respect to generative principles and observational material than the one that now commands our attention. To support inference in this extended universe of discourse about assessment, we will simply have to work through many problems from first principles. We must figure out just what it is we want to make inferences about—that is, first aspects, then models, of student competence. We must learn to construct situations that evoke evidence about these. We must explicate the probabilistic structure between the non–observable constructs and observations. We must (as is the focus of the present paper) use analytical methods that characterize the import and weight of evidence for our inferences. Sometimes this will be standard, familiar test theory, such as classical test theory, item response, or factor analysis. Sometimes it will not be. But probability–based inference can be gainfully applied to attack many of these problems, if not always with off–the–shelf tools. The first order of business for those of us in test theory, therefore, is to develop conceptual framework and analytic tools for carrying out these studies.

References

American Council on the Training of Foreign Languages. (1989). *ACTFL proficiency guidelines.* Yonkers, NY: Author.

Andersen, S.K., Jensen, F.V., Olesen, K.G., and Jensen, F. (1989). *HUGIN: A shell for building Bayesian belief universes for expert systems* [computer program] Aalborg, Denmark: HUGIN Expert Ltd.

Andreassen, S., Woldbye, M., Falck, B., and Andersen, S.K. (1987). MUNIN: A causal probabilistic network for interpretation of electromyographic findings. *Proceedings of the 10th International Joint Conference on Artificial Intelligence.* Milan: Kaufmann. pp. 366–372.

Doyle, A.C. (1930). *The complete works of Sherlock Holmes.* New York: Doubleday.

Estes, W.K. (1981). Intelligence and learning. In M.P. Friedman , J.P. Das, and N. O'Connor (Eds.), *Intelligence and learning.* New York: Plenum. pp. 3–23.

Glaser, R. (1991). Expertise and assessment. In M.C. Wittrock and E.L. Baker (Eds.), *Testing and cognition.* Englewood Cliffs, NJ: Prentice Hall. pp.17–30.

Greeno, J.G. (1989). A perspective on thinking. *American Psychologist, 44,* 134–141.

Haertel, E.H. (1989). Using restricted latent class models to map the skill structure of achievement test items. *Journal of Educational Measurement, 26,* 301–321.

Haertel, E.H., and Wiley, D.E. (1993). Representations of ability structures: Implications for testing. In N. Frederiksen, R.J. Mislevy, and I.I. Bejar (Eds.), *Test theory for a new generation of tests.* Hillsdale, NJ: Erlbaum. pp. 359–384.

Kahneman, D., Slovic, P., and Tversky, A. (1982). Judgment under uncertainty: Heuristics and biases. Cambridge: Cambridge University Press.

Krathwohl, D.R., and Payne, D.A. (1971). Defining and assessing educational objectives. In R.L. Thorndike (Ed.), *Educational measurement.* (2nd Ed.) Washington, D.C.: American Council on Education. pp. 17–45.

Kuhn, T.S. (1970). *The structure of scientific revolutions* (2nd edition). Chicago: University of Chicago Press.

Lauritzen, S.L., and Spiegelhalter, D.J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society, Series B, 50,* 157–224.

Lesh, R.A., and Lamon, S. (Eds.) (1992). *Assessments of authentic performance in school mathematics.* Washington, DC: American Association for the Advancement of Science.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd Ed.) New York: American Council on Education/Macmillan. pp. 13–103.

Mislevy, R.J. (1995). Probability–based inference in cognitive diagnosis. In P. Nichols, S. Chipman, and R. Brennan (Eds.), *Cognitively diagnostic assessment.* Hillsdale, NJ: Erlbaum.

Mislevy, R.J. (1994). Evidence and inference in educational assessment. 1994 Presidential address to the Psychometric society. *Psychometrika.*

Myford, C.M., and Mislevy, R.J. (1995). Monitoring and improving a portfolio assessment system. Center for Performance Assessment Research Report. Princeton, NJ: Center for Performance Assessment, Educational Testing Service.

Platt, W.J. (1975). Policy making and international studies in educational evaluation. In A.C. Purves and D.U. Levine (Eds.), *Educational policy and international assessment.* Berkeley, CA: McCutchen. pp. 33–59.

Noetic Systems, Inc. (1991). *ERGO* [computer program]. Baltimore, MD: Author.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* San Mateo, CA: Kaufmann.

Resnick, L.B., and Resnick, D.P. (1989). Assessing the thinking curriculum: New tools for educational reform. In B.R. Gifford and M.C. O'Conner (Eds.), *Future assessments: Changing views of aptitude, achievement, and instruction* Boston: Kluwer Publishers. pp. 37–75.

Schum, D.A. (1987). *Evidence and inference for the intelligence analyst.* Lanham, Md.: University Press of America.

Shafer, G., and Shenoy, P. (1988). Bayesian and belief–function propagation. *Working Paper 121.* Lawrence, KS: School of Business, University of Kansas.

Shavelson, R.J., Baxter, G.P., and Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher, 21*(4), 22–27.

Stake, R.E. (1991). The teacher, standardized testing, and prospects of revolution. *Phi Delta Kappan, 73,* 243–247.

Table 1: Excerpts from the ACTFL Proficiency Guidelines for Reading*

Level	Generic Description
Novice–Low	Able occasionally to identify isolated words and/or major phrases when strongly supported by context.
Intermediate–Mid	Able to read consistently with increased understanding simple connected texts dealing with a variety of basic and social needs. They impart basic information about which the reader has to make minimal suppositions and **to which the reader brings personal information and/or knowledge**. Examples may include short, straightforward descriptions of persons, places, and things, written for a wide audience. [emphasis added]
Advanced	Able to read somewhat longer prose of several paragraphs in length, particularly if presented with a clear underlying structure. **Comprehension derives not only from situational and subject matter knowledge but from increasing control of the language.** Texts at this level include descriptions and narrations such as simple short stories, news items, bibliographical information, social notices, personal correspondence, routinized business letters, and simple technical material written for the general reader. [emphasis added]
Advanced–Plus	Able to understand parts of texts which are conceptually abstract and linguistically complex, and/or **texts which treat unfamiliar topics and situations**, as well as some texts which involve aspects of target–language culture. Able to comprehend the facts to make appropriate inferences. [emphasis added]
Superior	Able to read with almost complete comprehension and at normal speed expository prose on **unfamiliar subjects** and a variety of literary texts. Reading ability is not dependent on subject matter knowledge, although the reader is not expected to comprehend thoroughly texts which are highly dependent on the knowledge of the target culture. At the superior level the reader can match strategies, top–down or bottom–up, which are most appropriate to the text.

*. Based on the *ACTFL proficiency guidelines*, American Council on the Training of Foreign Languages (1989).

Figure 1: Conditional Probabilities of extended-performance task responses, given competence level (deductive reasoning: three ACTFL levels, four levels of performance)

Note: Nodes represent variables. Bars represent probabilities of potential values of a variable, adding up to one. A dashed bar represents certainty.

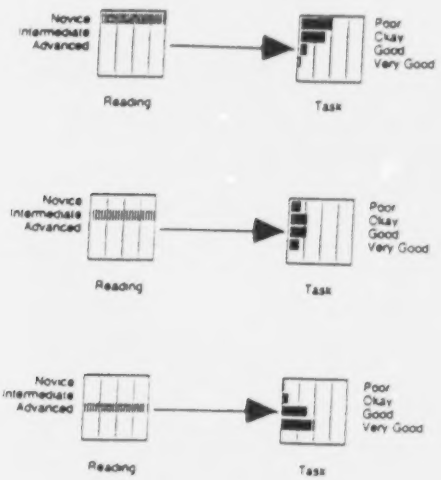
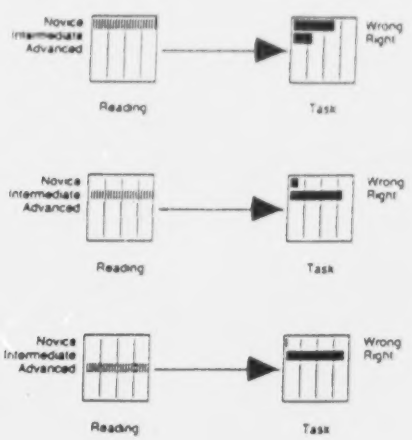


Figure 2: Conditional Probabilities of multiple-choice task responses, given competence level (deductive reasoning: three ACTFL levels, right/wrong performance)

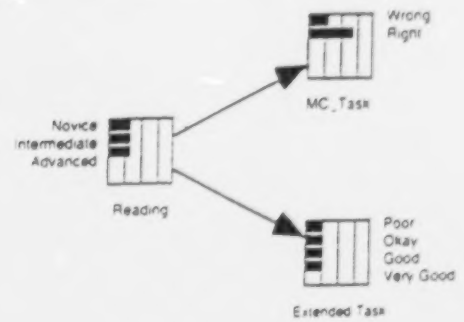
Note: Nodes represent variables. Bars represent probabilities of potential values of a variable, adding up to one. A dashed bar represents certainty.

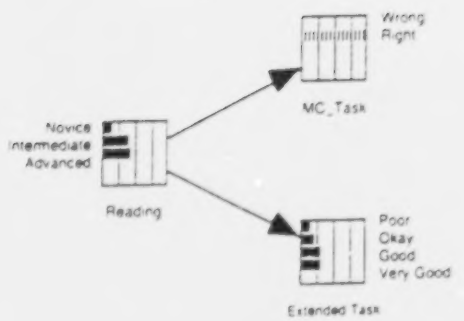


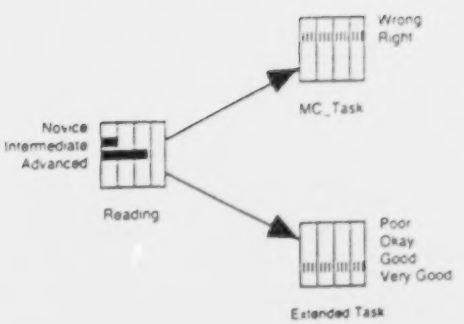
BEST COPY AVAILABLE

Figure 5: Successive updating of belief about competence level, after observing multiple-choice, then extended performance, task results: a) belief prior to observing any responses; b) belief after observing a correct multiple-choice response; c) belief after observing a correct multiple-choice response and a 'very good' extended-performance response.

Note: Nodes represent variables. Bars represent probabilities of potential values of a variable, adding up to one. A dashed bar represents certainty.



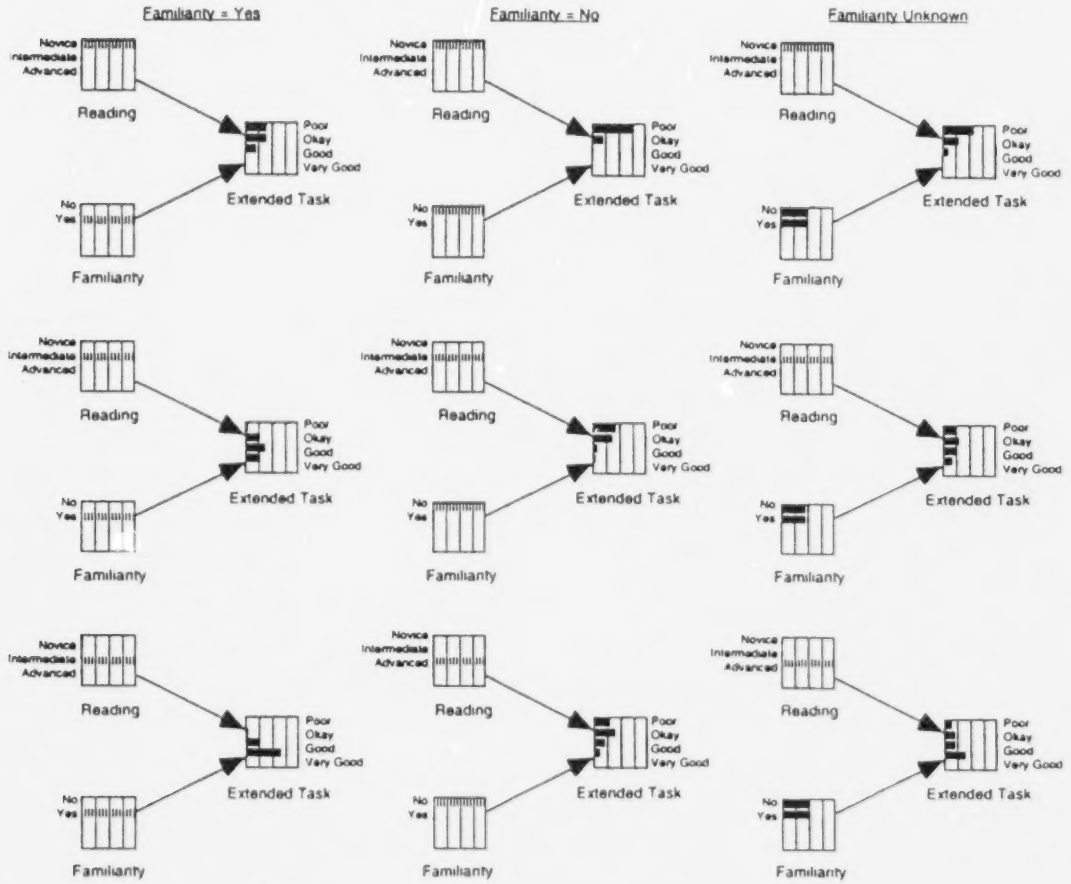




BEST COPY AVAILABLE

Figure 6: Conditional Probabilities of extended-performance, given competence level and task familiarity

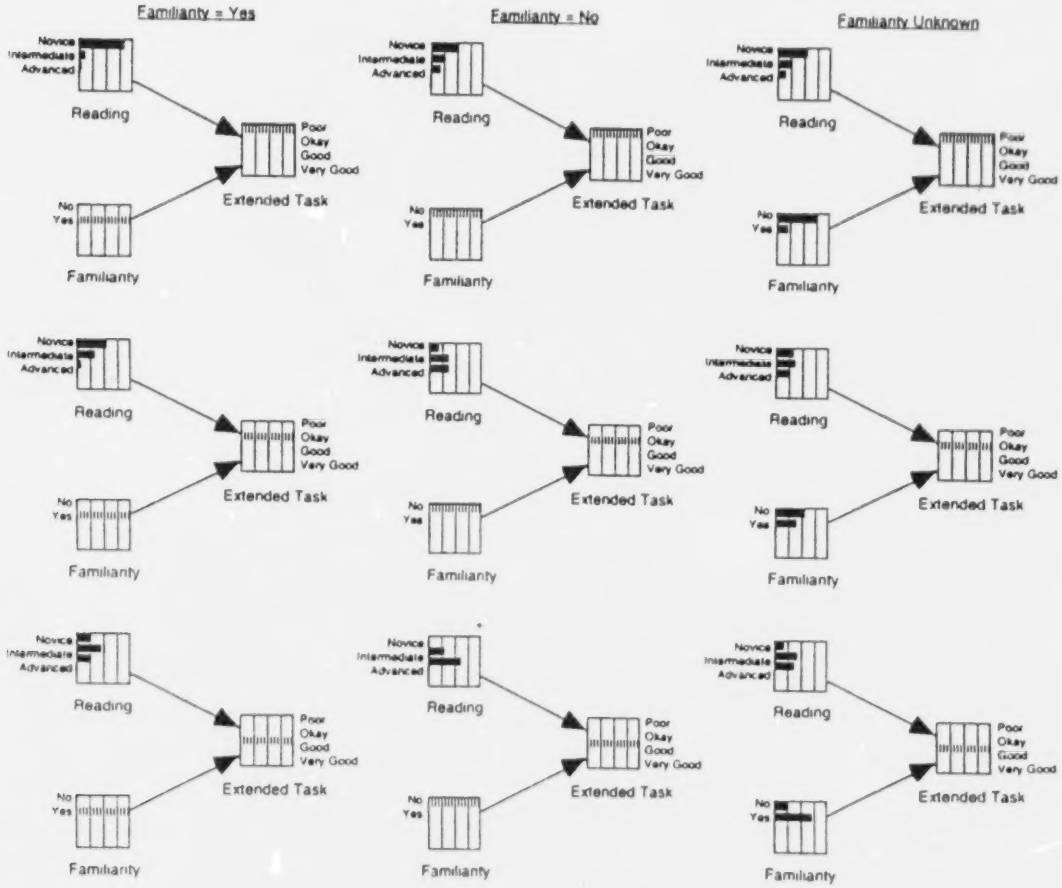
Note: Nodes represent variables. Bars represent probabilities of potential values of a variable, adding up to one. A dashed bar represents certainty.



BEST COPY AVAILABLE

Figure 7: Posterior Probabilities of competence level, given extended-task performance and task familiarity

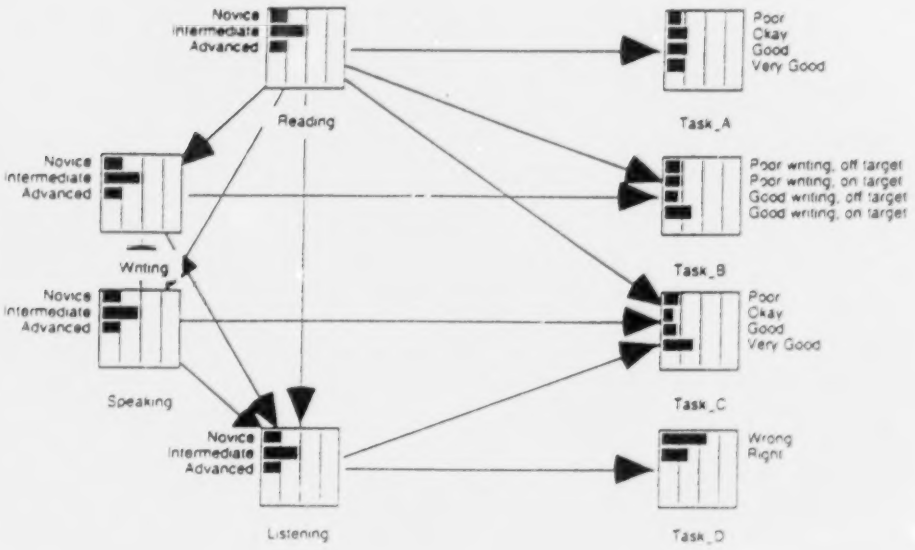
Note: Nodes represent variables. Bars represent probabilities of potential values of a variable, adding up to one. A dashed bar represents certainty.



BEST COPY AVAILABLE

Figure 8: Evidential structure of four tasks and four aspects of competence (status of belief before observing any responses)

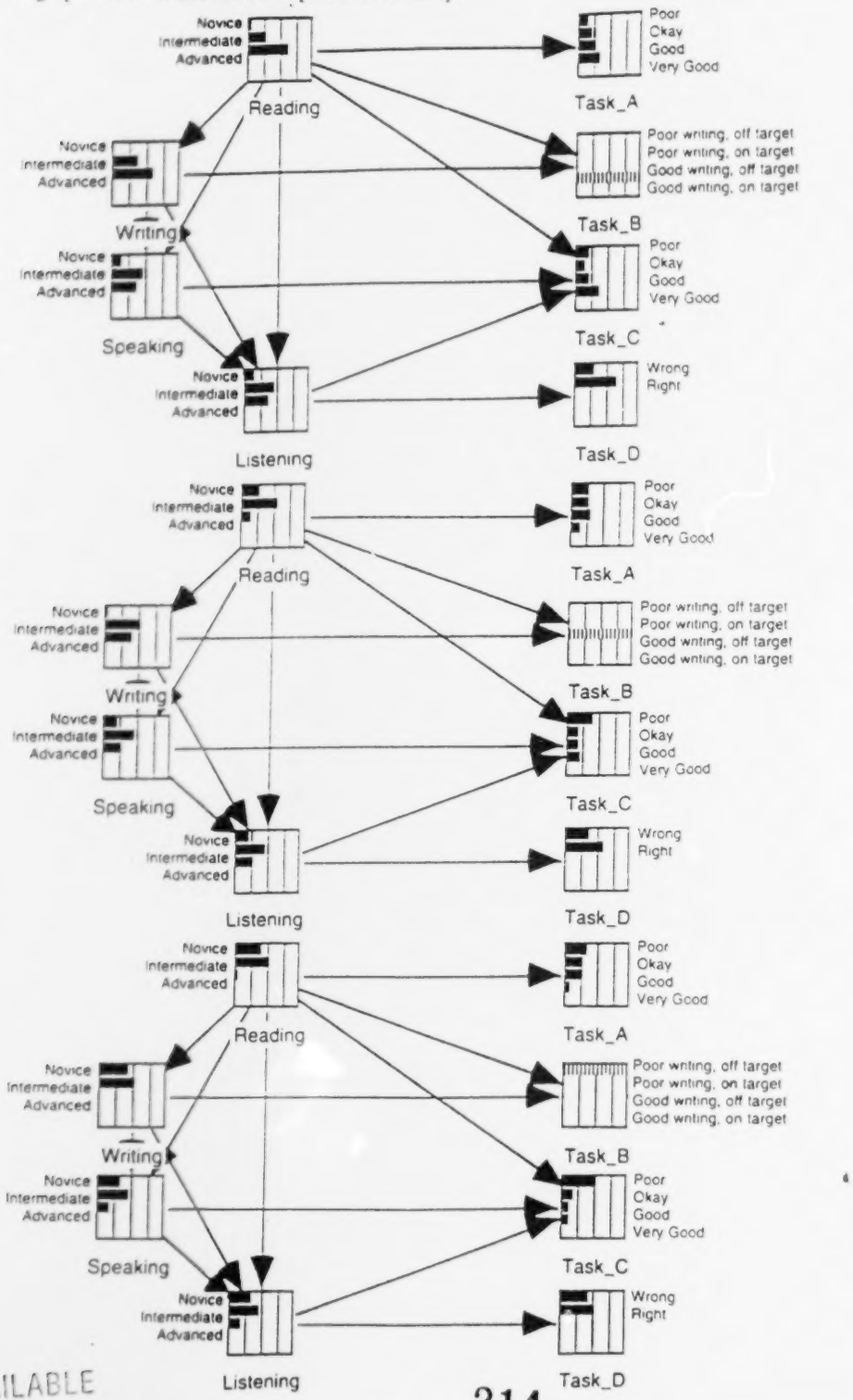
Note: Nodes represent variables. Bars represent probabilities of potential values of a variable, adding up to one. A dashed bar represents certainty.



BEST COPY AVAILABLE

Figure 9: Posterior Probabilities for competences, after observing various Task B responses

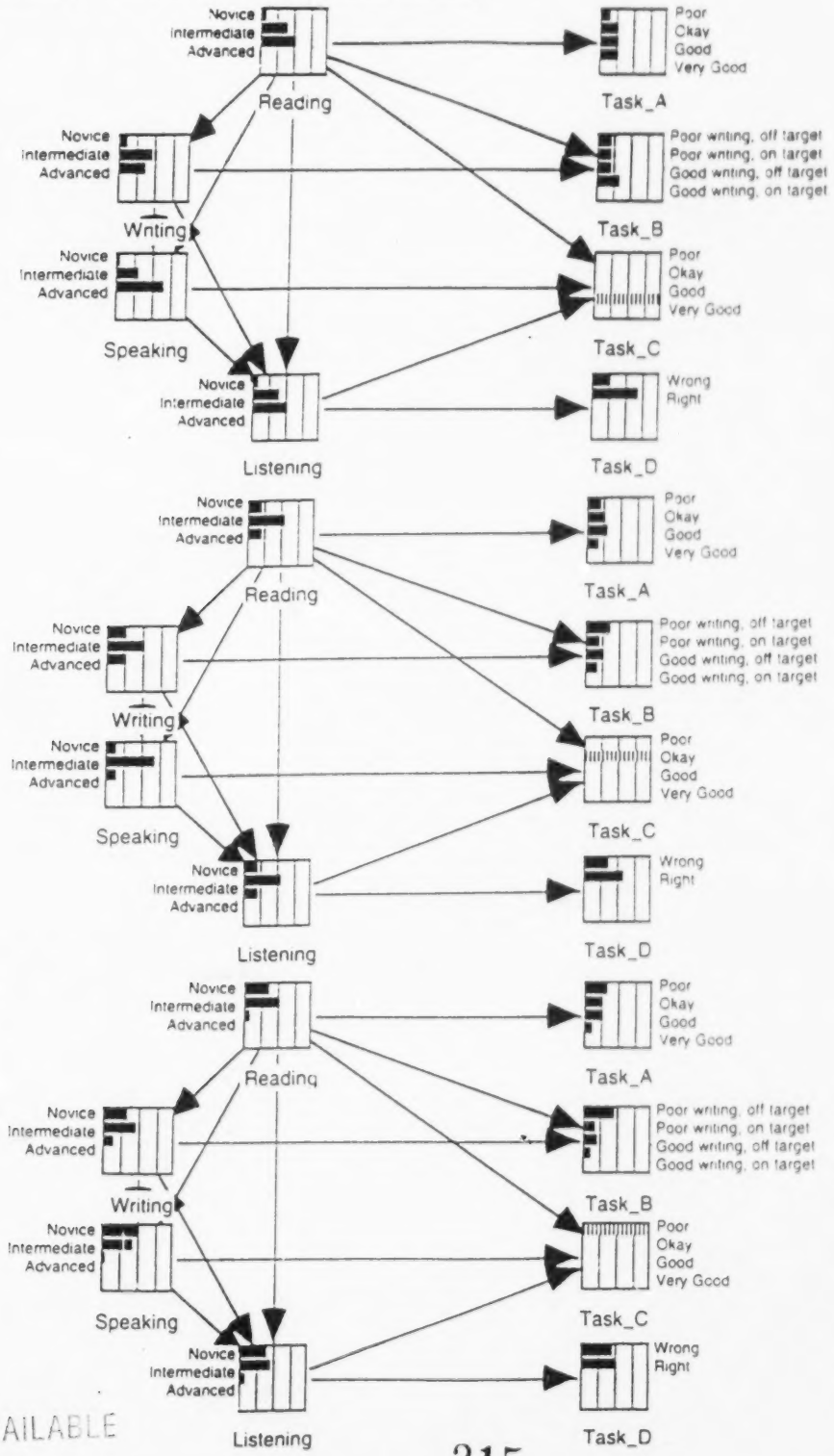
Note: Nodes represent variables. Bars represent probabilities of potential values of a variable, adding up to one. A dashed bar represents certainty.



BEST COPY AVAILABLE

Figure 10: Posterior Probabilities for competences, after observing various Task C responses

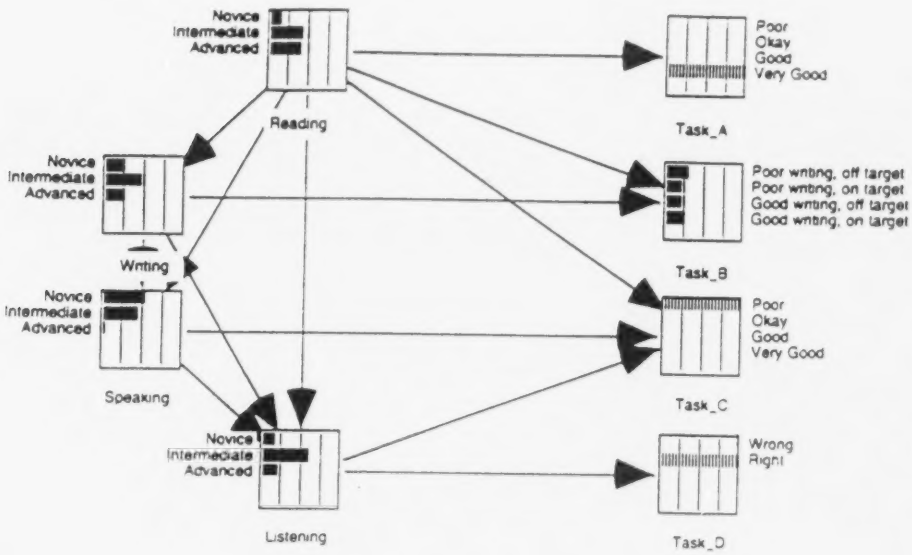
Note: Nodes represent variables. Bars represent probabilities of potential values of a variable, adding up to one. A dashed bar represents certainty.



BEST COPY AVAILABLE

Figure 11: Posterior Probabilities for competences, after observing a poor task C response, a very good Task A response, and a correct Task D response

Note: Nodes represent variables. Bars represent probabilities of potential values of a variable, adding up to one. A dashed bar represents certainty.



BEST COPY AVAILABLE

Appendix A: LAIS Program

SUNDAY, SEPTEMBER 25, 1994

Opening Session
Welcome: Betty Kilgore, Director / CALL
Introduction to Symposium: Eduardo C. Cascallar, Symposium Chair / Testing & Research Coordinator / CALL
Invited Speaker: Dr. Bernard Spolsky, Bar-Ilan University, Israel

MONDAY, SEPTEMBER 26, 1994

Roundtable Session
Title: You, the Government, and Language Aptitude
Introduction: E. Cascallar, Coordinator / CALL
Participants: Pardee Lowe, Jr. (CIA) (Chair), Marijke I. Cascallar (FBI), James R. Child (NCS), Madeline E. Ehrman (FSI), Danielle Janczewski (CIA), and John A. Lett, Jr. (DLI)

Plenary
Invited Speaker: Dr. Robert Sternberg, Yale University

Paper Session I
Robert N. Bostrom, Current Research in Measuring "Listening"
Madeline E. Ehrman, Is the Modern Language Aptitude Test Still Useful for Communicative Language Teaching?
Helen Lunt, The Investigation of Oral Proficiency and Language Learning Strategies in a Migrant ESL Context
Christine A. Montgomery, Effecting Changes in Affective Factors

Paper Session II
John A. Lett and John W. Thain, The Defense Language Aptitude Battery: What Is It and How Well Does It Work?
Madeline E. Ehrman, Expanding the Definition of Language Aptitude: The Role of Personality Variables
Pardee Lowe, Jr., Zero-Based Language Aptitude Test Design or Where's the Test's Focus?
James Child, Aptitude Tests: Conception and Design

Plenary
Invited Speaker: Dr. John de Jong, CiTO, The Netherlands

TUESDAY, SEPTEMBER 27, 1994

Plenary
Invited Speaker: Dr. Barry McLaughlin, University of California, Santa Cruz

Paper Session III
J. M. O'Malley and Anna Uhl Chamot, Learner Characteristics in Second Language Acquisition
Francis E. O'Mara and John W. Thain, Improving the Measurement of Language Aptitude: A Psychometric Analysis of the Defense Language Aptitude Battery

Paper Session IV
John W. Thain and John A. Lett, Improving the Measurement of Language Aptitude: The Potential Contribution of L1 Measures
Landes Holbrook, Eric Ott, Mary Lee Scott, and Cheryl Brown, A Factor Analytic Study of Language Learning Strategy Use by Older and Younger Adults

Workshop
Madeline Ehrman, Exploring Your Own Learning Style

Paper Session V
Brian MacWhinney, Psycholinguistic Issues in the Assessment of the Sub-Components of Language Abilities
Frank Borchardt, Ellis Page, and Fred Jacome, Let Computers Use the Past to Predict the Future: Using Machine-Based Retrospective Correlation Data for Prospective Aptitude Assessment Useful for Communicative Language Teaching?

Discussion Group
Title: Applications and Impact of Language Aptitude Assessment: Theoretical, Ethical, and Practical Issues
Facilitator: Eduardo Cascallar with Invited Speakers and Other Presenters

Plenary
Invited Speaker: Dr. Robert Mislevy, Educational Testing Service